大数据财务分析

DASHUJU CAIWU FENXI

策划编辑：袁相芬
责任编辑：张海红
封面设计：刘文东

高等职业教育大数据与会计系列教材

大数据财务分析

主编 黄雪雁 黄兰君 赵冉

DASHUJU CAIWU FENXI

北京邮电大学出版社

定价：49.80元

高等职业教育大数据与会计系列教材

“互联网＋”创新型教材

大数据财务分析

主编 黄雪雁 黄兰君 赵冉

北京邮电大学出版社
www.buptpress.com

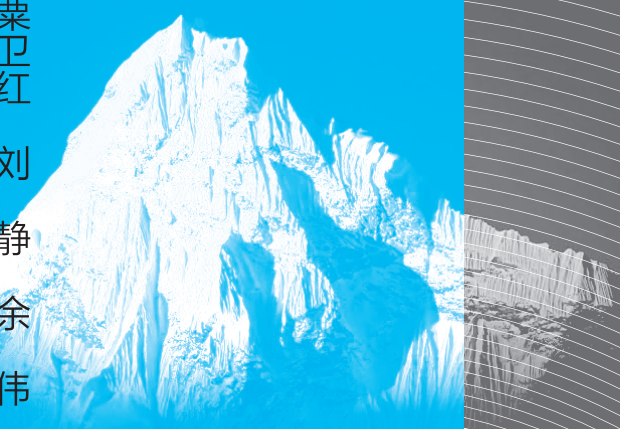
高等职业教育大数据与会计系列教材

“互联网 + ”创新型教材

大数据财务分析

主编 黄雪雁 黄三君 赵冉

副主编 粟卫红 刘静 余伟



北京邮电大学出版社
www. buptpress. com

内 容 简 介

本书根据当前财务大数据发展的新形势以及新需要编写，共包括七个项目，分别是大数据概述、数据处理与准备、资产负债表分析、利润表分析、现金流量表分析、财务比率分析、杜邦财务分析体系与经营业绩综合评价。

本书既可作为高职院校会计类专业的教学用书，也可作为会计工作者的参考用书。

图书在版编目(CIP)数据

大数据财务分析 / 黄雪雁，黄兰君，赵冉主编. -- 北京：北京邮电大学出版社，2023.8(2024.6 重印)
ISBN 978-7-5635-7018-8

Ⅰ. ①大… Ⅱ. ①黄… ②黄… ③赵… Ⅲ. ①财务管理—数据处理 Ⅳ. ①F275

中国国家版本馆 CIP 数据核字(2023)第 161810 号

策划编辑：袁相芬　　**责任编辑：**张海红　　**封面设计：**刘文东

出版发行：北京邮电大学出版社
社　　址：北京市海淀区西土城路 10 号
邮政编码：100876
发 行 部：电话：010-62282185　传真：010-62283578
E-mail：publish@bupt.edu.cn
经　　销：各地新华书店
印　　刷：三河市龙大印装有限公司
开　　本：787 mm×1 092 mm　1/16
印　　张：16.5 插页 1
字　　数：341 千字
版　　次：2023 年 8 月第 1 版
印　　次：2024 年 6 月第 2 次印刷

ISBN 978-7-5635-7018-8　　定　价：49.80 元

前言 PREFACE

大数据、信息技术、物联网、智能化的迅速发展，极大地改变了财务分析的外部环境和内在条件，使得财务分析的数据来源更加广泛，内容更加丰富。大数据财务分析在财务领域发挥着重要作用，具有很多优点：能实现数据快速传递和多维展示；分析工具丰富多样，实时可交互；分析结果可视化。随着大数据技术的不断变革，一方面，财务分析日益深入业务层面和经营层面，财务分析与业务分析、经营活动分析更加融合；另一方面，大数据分析技术、各种智能化管理软件大量应用于财务分析之中，其应用领域有不断深化的趋势，使财务分析呈现出许多新的特点和新的发展趋势。

基于对财务分析的上述理解，本书在体系内容的安排上，突出了以下几个特点。

（1）校企双元合作共同开发。本书由广西工商职业技术学院与北京首冠教育科技集团有限公司、厦门网中网软件有限公司合作编写，深度对接行业、企业标准，分析出岗位工作能力和典型工作任务，实践技能贴近行业需求并强化实操训练，使学生与未来的岗位要求相契合，满足社会的人才需求。

（2）融入课程思政元素。本书严格落实立德树人的根本任务，将每个项目的内容与课程思政案例合理融合，让学生在学习中了解国家政策，加强民族意识，增强职业道德感等，实现知识技能传授与价值引领的同频共振，引导学生正确的价值取向。

（3）以职业技能等级标准为依据，实现岗课赛证融通。本书以职业技能等级标准为依据来设计框架，依托首冠云实训平台，按大数据处理流程将数据采集、加工、分析与挖掘、数据可视化等最新技术融入其中；对接“1+X”大数据财务分析职业技能等级证书及职业技能大赛新设赛项的要求来精心设计和组织内容。

（4）融入数字资源。本书依托资源库平台建设在线课程及相关操作视频、微课、PPT 课件、实训操作等资源，并精选部分内容以数字资源的形式体现，学生通过移动终端扫描二维码即可进行学习，以实现线上线下学习的自由转换。

本书各项目的学时分配建议如下表所示。

教学内容	建议学时
项目一　认识大数据	3
项目二　数据处理与准备	4
项目三　资产负债表分析	5
项目四　利润表分析	5
项目五　现金流量表分析	6
项目六　财务比率分析	5
项目七　杜邦财务分析体系与经营业绩综合评价	5
总学时	33

本书由广西工商职业技术学院黄雪雁、黄兰君、赵冉任主编，由广西工商职业技术学院粟卫红、刘静、余伟任副主编，广西工商职业技术学院潘丽佳、廖晶、莫丹苗、陈锦齐、邓晨霞、蒋文利，北京首冠教育科技集团有限公司李建波、刘俊涛，厦门网中网软件有限公司胡世杰参与编写。

由于编者水平有限，书中难免存在不足之处，恳请广大读者批评指正。

编　者

目录 CONTENTS

项目二　数据处理与准备

知识目标

◎ 了解大数据的处理流程。

◎ 了解云计算的概念。

◎ 掌握大数据的分类知识。

◎ 能够运用 SQL 进行报表查询、连接。

素质目标

◎ 培养学生严谨的工作作风，进一步提高学生的职业素养。

◎ 提高学生运用数据思维和数据语言来分析数据的实践能力。

书证融通

大数据财务分析职业技能等级证书与本项目相关的考试内容及要求如下。

（1）了解大数据的处理流程。

（2）了解云计算的服务模式。

（3）能运用 SQL 结合财务专业技能进行数据分析，为企业管理者做决策提供依据。

任务一 了解大数据处理的基础架构

任务引入

在大数据时代背景下，我们应如何加工处理，才能使烦琐的数据成为决策力更强、更有效及多样化的信息资产呢？

必备知识

基于大数据的特征可以得知，通过传统 IT 技术存储和处理大数据的成本高昂。一家企业要大力发展大数据应用首先需要解决两个问题：一是低成本、快速地对海量、多类别的数据进行抽取和存储；二是使用新的技术对数据进行分析和挖掘，为企业创造价值。因此，大数据的存储和处理与云计算的技术密不可分。在当前的技术条件下，基于廉价硬件的分布式系统（如 Hadoop 技术等）被认为是最适合处理大数据的技术平台。

大数据处理流程主要包含数据收集、数据清洗和预处理、数据存储、数据剖析、数据可视化等环节。大数据质量贯穿于整个大数据处理流程，每一个数据处理环节都会对大数据质量产生影响。

一、数据收集

数据收集是指运用多个数据库来接收发自客户端（Web、App 或者传感器方式等）的数据，而且用户能够通过这些数据库进行简略的查询和处理作业。在数据收集的过程中，其主要特色是并发数高，因为可能同时会有成千上万的用户到访和操作。

二、数据清洗和预处理

虽然收集端本身会有许多数据库，但是要对这些海量数据进行有效的剖析，还应该将这些来自前端的数据导入一个集中的大型分布式数据库或分布式存储集群，而且能够在导入的基础上做一些简略的清洗和预处理作业。数据清洗和预处理进程的特色主要是导入的数据量大，每秒的导入量经常会达到百兆，甚至千兆等级。

三、数据存储

数据存储是指将记录和数字信息保存在磁性、光学或机械介质上，以供当前或未来运营之用。随着互联网的发展，越来越多的信息被数字化。数据呈爆炸式增长，数据的存储需求也越来越大，数据的多样化以及对重要数据的保护都对数据管理有了更高的要求，存

储产品不再是连接到服务器的辅助设备，而是成为互联网的主要成本。

四、数据剖析

数据剖析是指主要运用分布式数据库或分布式核算集群对存储于其内的海量数据进行普通的剖析和分类汇总等，以满足大多数常见的剖析需求。在这方面，一些实时性数据的剖析会用到 EMC 的 GreenPlum、Oracle 的 Exadata，以及根据 MySQL 的列式存储 Infobright 等，而一些批处理或半结构化的数据的剖析需求能够运用 Hadoop 技术。数据剖析的主要特色是触及的数据量大，其对系统资源，特别是对输入 / 输出（Input/Output，I/O）会有极大的占用。

五、数据可视化

数据可视化是指将数据以图形、图像的形式表示，并利用数据分析和开发工具发现其中未知信息的处理过程。数据可视化的目的是以清晰且高效的方式将信息传递给用户，并利用人眼的感知能力对数据进行交互的可视化表达，以增强用户对数据的认知。

任务二 了解云计算

任务引入

你知道什么是云计算吗？云计算对大数据的发展有什么作用？

必备知识

云计算是一种全新的领先信息技术，它能够结合 IT 技术和互联网实现超级计算和存储。随着 5G 互联网的发展与加持，各大互联网公司先后入局“云计算”领域，各自发展战略布局，如阿里云、华为云、腾讯云等。利用云计算的模型将信息、金融和服务等功能分散到由庞大分支机构构成的互联网“云”中，共享互联网资源，从而解决现实问题，并且达到高效率、低成本的目标。

一、云计算的概念及发展历程

（一）云计算的概念

云计算是指由位于网络上的一组服务器把其计算、存储、数据等资源以服务的形式提供给请求者以完成信息处理任务的方法和过程。在此过程中被服务者只是提出需求并获取

服务结果，对于需求被服务的过程并不知情。同时服务者以最优利用的方式动态地把资源分配给众多的服务请求者，以求达到最大效益。

云计算的核心思想是将大量用网络连接的计算资源进行统一管理和调度，构成一个计算资源池向用户按需服务。

（二）云计算的发展历程

“云”中的资源在使用者看来是可以无限扩展的，并且可以随时获取，按需使用；可以随时扩展，按使用付费。这种特性经常被称为像使用水电一样使用 IT 基础设施。

云计算的发展历程如图 2-1 所示。

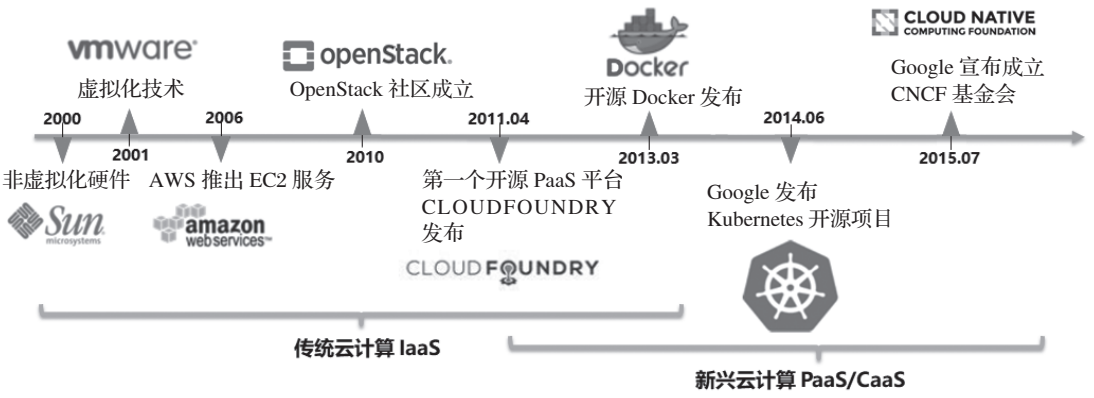


图 2-1

二、云计算的服务类别和类型

（一）云计算的服务类别

云计算有三种服务类别，分别为基础设施即服务（infrastructure as a service，IaaS）、平台即服务（platform as a service，PaaS）、软件即服务（software as a service，SaaS）。

1. 基础设施即服务

基础设施即服务是主要的服务类别之一，它将基础设施，包括处理能力、存储、网络和其他计算资源作为一种服务提供给用户使用，使后者可以在其上部署和运行包括操作系统和应用在内的任意软件。

2. 平台即服务

平台即服务也是一种服务类别。平台即服务是一个为开发人员提供通过全球互联网构建应用程序和服务的平台。平台即服务为开发、测试和管理软件应用程序提供按需开发环境。

3. 软件即服务

软件即服务，即云计算提供商通过互联网提供应用程序，并托管或管理应用程序，允许其用户通过互联网连接和访问应用程序，按需收费。

（二）云计算的类型

云计算按部署可以分为公有云、私有云和混合云。

1. 公有云

公有云面向所有用户提供服务。公有云是指云计算服务由第三方提供商完全承载和管理，为用户提供价格合理的计算资源访问服务，用户无须购买硬件、软件或基础架构，只需为其使用的资源付费。公有云用户无须支付硬件和带宽费用，投入成本低，但数据安全性低于私有云。

2. 私有云

私有云只为特定的用户提供服务，如大型企业出于安全考虑需要自己建设云环境，自己采购基础设施，搭建云平台，在此基础上开发应用云服务。私有云可充分保障虚拟化私有网络的安全，但投入成本比公有云高。

3. 混合云

混合云是公有云和私有云的结合，即将希望能被所有用户访问的数据放到公有云中，将需要安全保密的数据放到私有云中。混合云一般由用户创建，而管理和运维职责由用户和云计算提供商共同分担。混合云以私有云为基础的同时结合了公有云的服务策略，用户可根据业务私密性程度的不同，自主地在公有云和私有云之间进行切换。

三、云计算的特点

云计算是通过网络“云”使计算工作分布在大量的分布式计算机上，它的传输是通过互联网进行的。云计算是一种新型的计算模式，它用强大的计算能力来帮助人们解决各种各样的实际问题。例如，利用云计算可以提高生产效率、降低成本、节省能源，解决资源紧缺问题，技术共享问题等。目前被人们普遍接受的云计算的特点主要如下。

（一）高度虚拟化

高度虚拟化是指将服务器、存储器、网络等虚拟化。通过虚拟化技术将物理资源抽象整合，对资源进行动态分配和调度，从而在很大程度上减轻了数据中心的计算负担，同时也保障了数据安全。

（二）自动化

自动化包括对物理服务器、虚拟服务器的管理，对相关业务的自动化流程管理，对客户服务收费等的自动化管理，从而保证了运维的高效率。

（三）模块化

云计算集成了供配电制冷、机柜、气流遏制、综合布线动环监控等子系统，能够提高数据中心的整体运营效率，实现快速部署、弹性扩展和绿色节能。

（四）绿色节能

云计算数据中心无论是在硬件、软件上，还是在整体架构设计上，都强调绿色节能，

尤其 PUE 值（一种评价数据中心能源效率的指标）。和传统数据中心相比，云计算数据中心要求基础设施具有良好的弹性、扩展性、自动化、数据移动、多租户、空间效率和对虚拟化的业务。在每个约化程度方面，云计算数据中心在平台的运行效率更高，它支持多租户业务，能够针对每个租户的业务实现快速配置和部署。

四、云计算的应用

随着云计算技术的不断发展和它在存储、医疗、金融、教育等领域的应用不断深化，其对促进产业的发展起到了关键性的作用。

（一）存储云

存储云又称云存储，是在云计算技术的基础上发展起来的一种新的存储技术。存储云是一个以数据存储和管理为核心的云计算系统。用户可以将本地的资源上传至云端，可以在任何地方连入互联网来获取云上的资源。大家所熟知的谷歌、微软等大型网络公司就有云存储的服务。在国内，百度云和微云则是市场占有量很大的存储云。存储云向用户提供了存储容器服务、备份服务、归档服务和记录管理服务等，极大地方便了使用者对资源的管理。

（二）医疗云

医疗云是指在云计算、移动技术、多媒体、5G 通信、大数据及物联网等新技术基础上，结合医疗技术，使用云计算来创建医疗健康服务云平台，实现医疗资源的共享和医疗范围的扩大。云计算与医疗领域相结合提高了医疗机构的效率，使居民就医更加方便。医院的预约挂号、电子病历和电子医保卡等都是云计算与医疗领域相结合的产物。医疗云还具有数据安全、信息共享、动态扩展和布局全国的优势。

（三）金融云

金融云是指利用云计算的模型，将信息、金融和服务等功能分散到由庞大分支机构构成的互联网“云”中，旨在为银行、保险和基金等金融机构提供互联网处理和运行服务，同时共享互联网资源，从而解决现有问题并且达到高效率、低成本的目标。由于金融与云计算的结合，所以现在只需要在手机上简单操作，就可以完成银行存款、保险购买和基金买卖等业务。

（四）教育云

教育云实质上是指教育信息化的一种发展。具体来说，教育云可以将所需要的任何教育硬件资源虚拟化，然后将其传入互联网中，从而为使用者提供一个方便快捷的平台。例如，大规模开放的在线课程慕课（massive open online course，MOOC）就是教育云的一种应用。

任务三 掌握数据格式相关知识

任务引入

生活中常见的数据有哪些？它们分别属于什么格式的数据？

必备知识

数据格式（data format）是描述数据保存在文件或记录中的编排格式。数据格式可以是字符形式的文本格式，也可以是二进制数据形式的压缩格式。字符形式的文本格式占用的存储空间多但透明度高，二进制数据形式的压缩格式占用的存储空间少但透明度低。数据格式主要由数据类型及数据长度来描述。

一、数据类型

（一）按结构属性分类

按结构属性的不同，数据可以分为结构化数据与非结构化数据。它们不仅在存储形式上有所不同，在数据处理和数据分析的方法上也大相径庭。

1. 结构化数据

结构化数据是指可以使用二维表结构来表示和存储的数据。例如，Excel 数据、企业 ERP 系统数据、企业会计信息系统数据、银行交易记录数据等，它们大多存储在大型数据库中，方便用户进行检索、分析和处理。

2. 非结构化数据

非结构化数据是指不能用二维表结构来表示和存储的数据。非结构化数据没有统一的规则，涉及音（视）频、图片、文本等形式。例如，从网站抓取的新闻或者媒体数据、某个电视剧或者电影的评价数据等，都需要通过一定的方法将这些数据转化为结构化数据，再进行有效的分析。

（二）按连续特征分类

按连续特征的不同，数据可以分为连续型数据与离散型数据。

1. 连续型数据

连续型数据是指在一定区间内可以连续取值的数据，如人的体重数值、气温的度数、

电视剧的收视率等。

2. 离散型数据

离散型数据又称为不连续数据，其取值只能用自然数或整数表达，如硬币的正反面取值、某人的学历取值等。

（三）按测量尺度分类

按测量尺度的不同，数据可分为定类数据、定序数据、定距数据和定比数据。

1. 定类数据

定类数据的数据表现为类别，不能区分顺序，无法描述大小、高度、重量等信息；不能进行任何运算，主要用于标识数据所描述的主体对象的类别或者属性名称。例如，人的性别分为男性和女性两类，量化后可分别用数字 0 和 1 表示；企业按行业可分为旅游业、教育业、制造业等，量化后可分别用数字 1、2、3 表示。这些数字只是代号，不能进行任何数学运算。

2. 定序数据

定序数据的数据表现为类别，有顺序但不能进行计算，也称为序列数据，用于按顺序描述事物所具有的属性。定序数据虽然可以用数字或者序号来排列，但并不代表数据的大小，只代表数据之间的顺序关系。例如，人的受教育程度可分为高中、大学本科、硕士研究生、博士研究生，可分别用数字 1、2、3、4 表示，这些只代表顺序，按照大小正序排列。

3. 定距数据

定距数据也可以称为间隔尺度，相较定类数据和定序数据而言，它能对事物进行准确测量。定距数据不仅能比较各类事物的优劣，还能计算出事物之间差异的大小，所以其数据表现为数值。定距数据具有间距特征的变量，有单位，没有绝对零点，可以做加减运算，但不能做乘除运算，如温度。

4. 定比数据

定比数据是由定比尺度计量形成的，其数据表现为数值，既可以进行加减运算，也可以进行乘除运算。定比数据代表数据的最高级，既有测量单位，也有绝对零点（可以取值为 0）。例如，小明的体重是 60 kg，小刚的体重是 30 kg，我们可以说小明的体重是小刚体重的 2 倍。

由此可以看出，定类数据和定序数据的数据表现为类别，属于定性数据；定距数据和定比数据的数据表现为数值，属于定量数据。

二、数据长度

数据长度是指数据占用多少个字节。数据占用的字节越多，能存储的数据就越多，对于数字来说，值就会更大；反之，能存储的数据就有限。

任务四 了解数据库相关知识

任务引入

大数据存储的特点有哪些？大数据存储和传统数据存储有何不同？

必备知识

一、数据库的类型

数据库是电子化信息的集合，是指将信息规范化并使之电子化，形成电子信息“库”，以便利用计算机对这些信息进行快速有效地存储、检索、统计与管理。数据管理不再仅仅是存储和管理数据，而是转变成用户所需要的各种管理数据的方式。数据库有很多种类型，如从最简单的存储数据表格到能够进行海量数据存储的大型数据库系统，都在被广泛应用。

数据库技术是管理信息系统、办公自动化系统、决策支持系统等各类企业信息系统的核心部分，是进行科学研究和管理的重要技术手段。数据库可分为关系型数据库和非关系型数据库。

（一）关系型数据库

关系型数据库是指采用了关系模型来组织数据的数据库。其以行和列的形式存储数据，以便于用户理解。关系型数据库中一系列的行和列被称为表，一组表组成了数据库。

1. 关系型数据库的应用

在关系型数据库中，存储的格式可以直观地反映实体间的关系。关系型数据库和常见的表格比较相似。在关系型数据库中，表与表之间是有很多复杂的关联关系的。常见的关系型数据库有 MySQL、SQLServer 等。在构建轻量或者小型的应用中，使用不同的关系型数据库对系统的性能影响不大；但是在构建大型应用时，则需要根据应用的业务需求和性能需求，选择合适的关系型数据库。采用标准语言 SQL（structured query language），常见的操作有查询、新增、更新、删除、求和、排序等。关系型数据库对于结构化数据的处理更合适，如学生成绩、地址等，这样的数据在一般情况下需要使用结构化的查询。

2. 关系型数据库的视图

关系型数据库的视图就是一种虚拟表，是通过 SQL 语句按逻辑结构将表中数据进行重新整理的，其内容与真实表类似，包含一系列带有名称的列和行数据，但并不以存储数据

值的形式存在，行和列数据来源于查询所引用基本表，视图为动态生成。关系型数据库视图的主要特点如下。

（1）视图的列可以来自不同的表，是表的抽象在逻辑意义上建立的新关系。

（2）视图是由基本表（实表）产生的表（虚表）。

（3）视图的建立和删除不影响基本表。

（4）对视图内容的更新（添加、删除、修改）直接影响基本表。

（5）当视图来自多个基本表时，不允许添加和删除数据。

（二）非关系型数据库

非关系型数据库又被称为 NoSQL（not only SQL），描述的是大量结构化数据存储方法的集合。NoSQL 数据库技术具有非常明显的应用优势，如数据库结构相对简单，在大数据量下的读写性能好，能满足随时存储自定义数据格式的需求，非常适用于大数据处理工作等。

关系型数据库与非关系型数据库的区别如表 2-1 所示。

表 2-1　关系型数据库与非关系型数据库的区别

区　别	关系型数据库	非关系型数据库
存储方式	传统的关系型数据库采用表格储存的方式，数据以行和列的方式进行存储，要读取和查询都十分方便	非关系型数据库不适合表格存储的方式，通常以数据集的方式，将大量的数据集中存储在一起，类似于键值对、图结构或者文档
存储结构	按照结构化的方法存储数据，整个数据表的可靠性和稳定性都比较高，如需修改数据表的结构就会十分困难	采用动态结构，对于数据类型和结构的改变非常适应，可以根据数据存储的需要灵活地改变数据库的结构
存储规范	数据按照最小关系表的形式进行存储	用平面数据集的方式集中存放
扩展方式	只具备纵向扩展能力	存储方式一定是分布式的，它可以采用横向的方式来扩展数据库

二、大数据存储的概念和方式

（一）大数据存储的概念

大数据存储是指将大规模的数据存储在分布式存储系统中，以便于管理和处理。大数据存储通常采用分布式存储技术，将数据分散存储在不同的节点上，并通过网络连接形成一个整体，以提高数据的可靠性和可扩展性。

（二）大数据存储的方式

大数据存储的方式是指对结构化、半结构化和非结构化海量数据进行存储和管理，轻型数据库无法满足对其存储以及复杂的数据挖掘和分析操作，因此，通常使用分布式文件系统、NewSQL 和 NoSQL 数据库、云数据库等存储方式。

（1）分布式文件系统。分布式文件系统是通过网络实现文件在多台主机上进行分布式存储的文件系统。它是将固定于某个地点的某个文件系统，扩展到任意多个地点 / 多个文

件系统，众多的节点组成一个文件系统网络，每个节点可以分布在不同的地点，通过网络进行节点间的通信和数据传输。人们在使用分布式文件系统时，无须关心数据是存储在哪个节点上，或者是从哪个节点处获取的，只需要像使用本地文件系统一样管理和存储文件系统中的数据即可。

（2）NewSQL 和 NoSQL 数据库。NoSQL 数据库指的是非关系型的数据库，是对不同于传统的关系型数据库的数据库管理系统的统称，用于超大规模数据的存储，灵活的数据模型可以很好地支持 Web 2.0 应用，具有强大的横向扩展能力等。NewSQL 是对各种新的可扩展 / 高性能数据库的简称，这类数据库不仅具有 NoSQL 对海量数据的存储管理能力，还保持了传统数据库支持 ACID 和 SQL 等特性。

（3）云数据库。云数据库通过互联网连接数据库服务，将数据存储在云中，通过云计算的方式进行管理。与传统数据库相比，云数据库能够提供更高效的数据管理服务，更安全的数据存储服务，以及更灵活的数据连接服务。

三、大数据存储与传统数据存储的差异

大数据存储对于大数据处理和分析起到了重要的作用。大数据存储和传统数据存储的差异如表 2-2 所示。

表 2-2　大数据存储与传统数据存储的差异

区　　别	大数据存储	传统数据存储
存储方式	分布式存储	集中式存储
数据规模	大规模的数据，可以处理数十亿条数据	小规模的数据，通常只能处理几百万条数据
数据类型	需要处理结构化数据	需要处理半结构化和非结构化数据
数据处理方式	事务处理或批处理方式，即一次处理一定量的数据	实时处理方式，即对数据进行实时处理和分析
数据安全性	采用复杂的安全技术，如数据分片、数据备份等技术	访问控制和加密技术保护

任务五　掌握 SQL 数据处理知识

任务引入

常用的 SQL 语句有哪些？

必备知识

一、SQL 的定义

SQL 是指结构化查询语言，是用于访问和处理数据库的标准计算机语言。

二、SQL 的范围

SQL 的范围包括数据获取、更新、删除和插入，数据库模式创建和修改，以及数据访问控制。

三、SQL 语句

SQL 语句主要分为两个部分：数据操作语言（data manipulation language，DML）和数据定义语言（data definition language，DDL）。

（一）数据操作语言

数据操作语言主要是数据库用于获取、更新、删除和插入数据的语法。

SQL 中常用的 DML 语句如表 2-3 所示。

表 2-3　SQL 中常用的 DML 语句

语　句	说　明
SELECT	从数据库表中获取数据
UPDATE	更新数据库表中的数据
DELETE	从数据库表中删除数据
INSERT INTO	向数据库表中插入数据

（二）数据定义语言

数据定义语言可以创建或删除表格，也可以定义索引（键），规定表之间的链接，以及施加表之间的约束，主要有创建新数据库、修改数据库、创建新表、变更（改变）数据库表、删除表、创建索引（搜索键）、删除索引等操作。

SQL 中常用的 DDL 语句如表 2-4 所示。

表 2-4　SQL 中常用的 DDL 语句

语　句	说　明
CREATE DATABASE	创建新数据库
ALTER DATABASE	修改数据库
CREATE TABLE	创建新表
ALTER TABLE	变更（改变）数据库表
DROP TABLE	删除表
CREATE INDEX	创建索引（搜索键）
DROP INDEX	删除索引

四、SQL 语言中的查询语句

SQL 提供了结构形式一致但功能多样化的查询语句 Select。Select 语句一般形式的语义为：从给出表名的表格中，查询出符合检索条件的内容，将新表格按列名及顺序进行显示。Select 语句的一般形式为

Select 列名称 [[, 列名称]...]
From 表名称
[Where 检索条件];

注意：

（1）上述语句中的 [ ] 可不要。

（2）SQL 语句可以单行或多行书写，以分号结尾。

（3）可使用空格和缩进来增强语句的可读性。

（4）SQL 语句不分大小写，但某些数据库系统要求在每条 SQL 命令的末端使用分号。

（5）SQL 语句使用单引号来环绕文本值（大部分数据库系统也接受双引号）。如果是数值，请不要使用引号。

Where 子句中的运算符如表 2-5 所示。

表 2-5 Where 子句中的运算符

运 算 符	说 明
=	等于
< > 或 ! =	不等于
>	大于
<	小于
> =	大于等于
< =	小于等于
BETWEEN	在某个范围内
LIKE	搜索匹配的字符串模式

（一）基本查询

查询一个表中所有字段的语法为

Select*from 表名称

查询一个表中指定列的语法为

Select 列 1, 列 2 from 表名称

【例 2-1】学生信息表如表 2-6 所示。请按要求写出 SQL 查询语句。

表 2-6　学生信息表

姓　名	性　别	班　级	学　号	年　龄
李云龙	男	2021 班	2053202101	19
赵刚	男	2022 班	2053202203	20
孔捷	男	2023 班	2053202308	18

【要求 1】查询学生信息表中的所有信息。

【答案 1】Select * from 学生信息表

【答案 2】Select 姓名，性别，班级，学号，年龄 from 学生信息表

查询后显示的结果如表 2-7 所示。

表 2-7　查询后显示的结果 1

姓　名	性　别	班　级	学　号	年　龄
李云龙	男	2021 班	2053202101	19
赵刚	男	2022 班	2053202203	20
孔捷	男	2023 班	2053202308	18

【要求 2】查询学生信息表中学生的姓名和班级。

【答案】Select 姓名，班级 from 学生信息表

查询后显示的结果如表 2-8 所示。

表 2-8　查询后显示的结果 2

姓　名	班　级
李云龙	2021 班
赵刚	2022 班
孔捷	2023 班

【要求 3】查询学生信息表中年龄大于等于 19 岁的学生的姓名、性别、班级。

【答案】Select 姓名，性别，班级 from 学生信息表 where 年龄 > =19

查询后显示的结果如表 2-9 所示。

表 2-9　查询后显示的结果 3

姓　名	性　别	班　级
李云龙	男	2021 班
赵刚	男	2022 班

【要求 4】将学生信息表命名为 a，并查询 a 表中的姓名、性别和班级。

【答案】Select a. 姓名，a. 性别，a. 班级 from 学生信息表 a

查询后显示的结果如表 2-10 所示。

表 2-10 查询后显示的结果 4

姓　名	性　别	班　级
李云龙	男	2021 班
赵刚	男	2022 班
孔捷	男	2023 班

为学生信息表命名的语法为

Select a. 列 1, a. 列 2 from 表名称 as a

其中，a 就是表名称的新表名（as 可以省略）。

给表重新命名的原因如下。

（1）表名比较长。

（2）需要在多个表中进行查询并把查询内容同时输出。

（3）需要进行表连接。

（二）根据条件计算值的查询

1. 一般计算

语法格式为

Select 计算公式 1 计算公式 1 的列命名，计算公式 2 计算公式 2 的列命名 from 表名称

【例 2-2】员工工资表如表 2-11 所示。在表 2-11 中增加一列“工资 / 万元”，并在新增加的一列中计算出相应的金额。

表 2-11 员工工资表

姓　名	性　别	部　门	工资 / 元
李云龙	男	人力资源部	50 000
赵刚	男	销售部	80 000
孔捷	男	财务部	60 000

【答案】Select *, 工资 /10000 工资 / 万元 from 员工工资表

计算结果如表 2-12 所示。

表 2-12 计算结果 1

姓　名	性　别	部　门	工资 / 元	工资 / 万元
李云龙	男	人力资源部	50 000	5
赵刚	男	销售部	80 000	8
孔捷	男	财务部	60 000	6

2. 根据条件计算值

（1）只有一个条件表达式的语法。其语法格式为

Select …,（case when 条件表达式 then 值 1 else 值 2 end）

表达式结果列命名 from 表名称

“case when 条件表达式 then 值 1 else 值 2 end”的语义为：如果满足条件表达式，则显示值 1，不满足则显示值 2，如果值是空白，可用‘ ’表示。

【例 2-3】在表 2-11 中增加一列“员工房租扣款”，其结果的计算条件为：如果应发工资＞ 50 000 元，则员工房租扣款 = 应发工资 *0.10；如果应发工资≤ 50 000 元，则员工房租扣款 = 3 000 元。

【答案】Select *,（case when 应发工资 > 50 000 then 应发工资 *0.10 else 3000 end）员工房租扣款 from 员工工资表

计算结果如表 2-13 所示。

表 2-13 计算结果 2

姓　名	性　别	部　门	应发工资 / 元	员工房租扣款
李云龙	男	人力资源部	50 000	3 000 元
赵刚	男	销售部	80 000	8 000 元
孔捷	男	财务部	60 000	6 000 元

（2）有两个条件表达式的语法。其语法格式为

Select…,

（case when 条件表达式 1 then 值 1

when 条件表达式 2 then 值 2,

…

else 值 end）列命名

from 表名称

【例 2-4】在表 2-11 中增加一列“员工房租扣款”，其金额的计算条件为：如果应发工资≥ 50 000 元，则员工房租扣款 = 应发工资 *0.10；如果应发工资＞ 70 000 元，则员工房租扣款 = 应发工资 *0.20；如果员工工资＜ 50 000 元，则不用交房租。

【答案】Select *,（case when 应发工资＞ =50000 then 应发工资 *0.10　when 应发工资＞ 70000 then 应发工资 *0.20 else‘ ’end）员工房租扣款 from 员工工资表

计算结果如表 2-14 所示。

表 2-14 计算结果 3

姓　名	性　别	部　门	应发工资 / 元	员工房租扣款
李云龙	男	人力资源部	50 000	5 000 元
赵刚	男	销售部	80 000	16 000 元
孔捷	男	财务部	60 000	6 000 元

（三）多表连接的查询

多表连接的查询的语法格式为

```
Select 列名 from 表 1 [inner/left/right] join 表 2 on 连接条件
```

对多表连接的查询的语法格式的含义解释如下。

语义：从表 1 中查询到相应的列名后，根据匹配条件与表 2 合并在一起。

inner join：内连接，列出两个表中都存在的记录。

left join：左连接，即使没有匹配条件也列出左表中的所有记录。

right join：右连接，即使没有匹配条件也列出右表中的所有记录。

【例 2-5】员工代码表如表 2-15 所示。将表 2-11 和表 2-15 根据工号进行连接，并让员工工资表的内容全部显示。

表 2-15　员工代码表

员工代码	姓　　名
001	李云龙
002	赵刚
003	孔捷

【答案】Select * from 员工工资表 left join 员工代码表 on 员工工资表 . 员工代码 = 员工代码表 . 员工代码

计算结果如表 2-16 所示。

表 2-16　计算结果 4

员工代码	性　　别	部　　门	工资 / 元	姓　　名
001	男	人力资源部	50 000	李云龙
002	男	销售部	80 000	赵刚
003	男	财务部	60 000	孔捷

实训任务

企业在激烈的市场竞争中为了扩大销售额，需要结合自身实际情况、行业特点来制定相应的销售政策。但在具体实施过程中，可能会因为多方面因素的影响，致使销售政策无法得到有效的落实，并出现了账款难以回收的现象，这时不仅会对企业资金流产生不利影响，而且会增加货款催收成本，严重阻碍企业的发展。因此，做好应收账款风险管理至关重要，其既能够有效控制应收账款风险，又能够确保企业的快速发展。

一、实训要求

用 SQL 对应收账款的数据进行分析管理。应收账款数据表如表 2-17 所示，客户表如

表 2-18 所示。

表 2-17　应收账款数据表 1

发票编号	发票日期	到期日	实际收款日期	客户 ID	销售渠道	发票金额

注：表中实际收款日期为 9999-12-31，代表暂未收款。

表 2-18　客户表 1

客户 ID	客户名称

二、实训内容

实训内容如表 2-19 所示。

表 2-19　实训内容 1

任务序号	任务内容
1	通过导入数据，建立应收账款数据表与客户表之间的关联
2	新增列“实际收款日期 2”，清洗连接后的应收账款数据表和客户表的收款日期，将“实际收款日期”为“9999-12-31”的值变为空值
3	新增列“是否应收账款”，判断每笔业务是否为应收账款。若发票日期 < =2018-01-01 和实际收款日期 > 2018-01-01 或者实际收款日期为空值，该笔业务可确定为应收账款，则“是否应收账款”的定义值为“应收账款”，其余情况为空值
4	新增列“欠款天数”并计算，设置条件为“到期日 > 2018-01-01”，实际收款日期 < = 到期日，值均为 0，到期日为 9999-12-31，值为 2018-01-01 与到期日之间的时间，其余均为实际收款日期与到期日之间的时间
5	在任务 4 的基础上新增列“账龄”，设置条件为“欠款天数 =0，值为未欠款；欠款天数 < =30，值为 1—30 天；欠款天数 < =60，值为 31—60 天；欠款天数 < =90，值为 61—90 天；欠款天数 > 90，值为大于 90 天”
6	在任务 5 的基础上，新增列“预计坏账率”，设置条件为“欠款天数 =0，坏账率为 0.05；欠款天数 < =30 天，坏账率为 0.1；欠款天数 < =60 天，坏账率为 0.2；欠款天数 < =90 天，坏账率为 0.3；欠款天数 > 90 天，坏账率为 0.5”

注：涉及函数 DATEDIFF（ ），该函数表示返回两个日期之间的时间。

三、实训步骤

（1）启动大数据分析实验室，在“我的项目”中单击“创建新项目”，如图 2-2 所示。

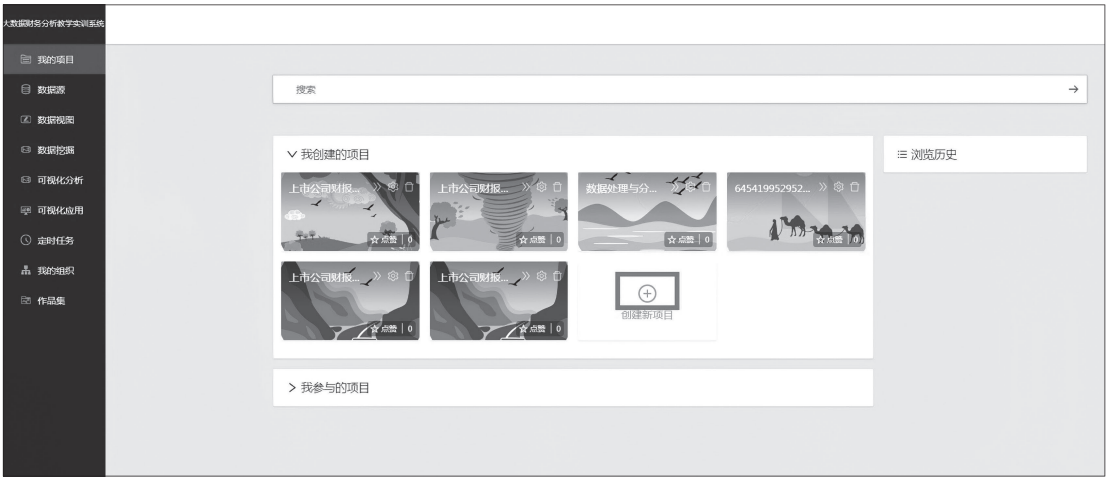


图 2-2

（2）弹出“创建项目”对话框，在“组织”文本框中选择“大数据与会计组织”，在“名称”文本框内输入“数据处理与分析 - 应收账款管理”，单击“保存”按钮，如图 2-3 所示。

图 2-3

（3）选择“我的项目”，单击“我创建的项目”中的“数据处理与分析 - 应收账款管理”，如图 2-4 所示。

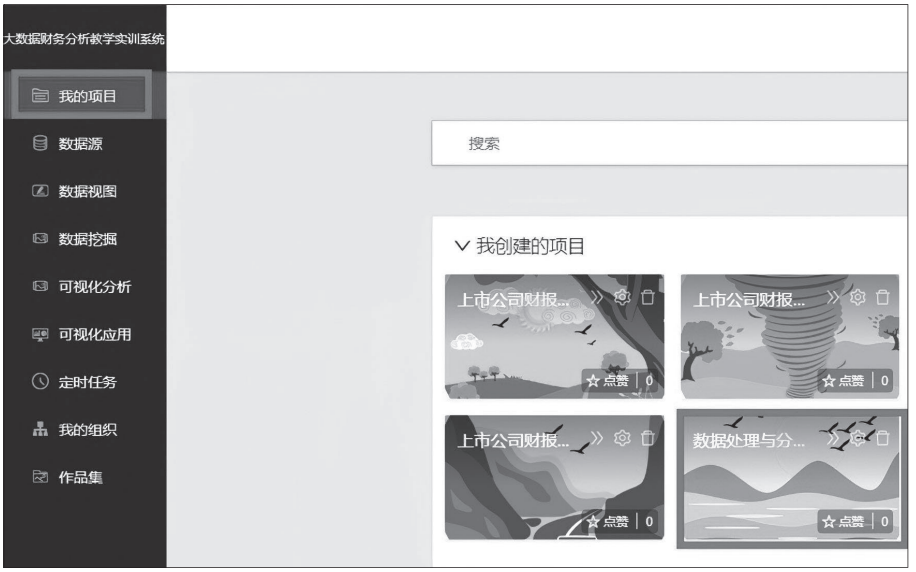


图 2-4

（4）进入项目后，选择“数据源”，单击右上角“挂载数据集”按钮，在弹出的“挂载数据集”对话框中选择“大数据财务数据集”，如图 2-5 所示。单击“确定”按钮，挂载成功。

（a）　　（b）　　（c）

图 2-5

（5）进入“数据视图”，找到该项目，在“操作”中单击修改按钮，如图 2-6 所示。

（a）　　（b）

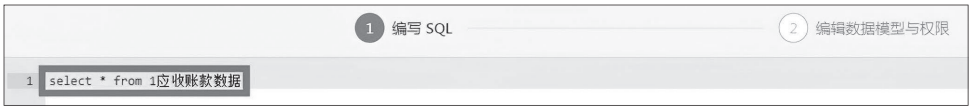
图 2-6

（6）进入数据编辑界面，选择数据源“大数据财务数据集”，如图 2-7 所示。用 SQL 语句查询表格“1 应收账款数据”，输入下列语句：

```
select * from 1 应收账款数据
```

单击右下角的“执行”按钮，如图 2-8 所示。

图 2-7



（a）

（b）

图 2-8

（7）连接应收账款数据表与客户表，输入下列语句：

select a. *, b. 客户名称 from 1 应收账款数据 a left join 8 客户表 b on a. 客户 ID=b. 客户 ID

然后单击右下角的“执行”按钮。执行成功后，如图 2-9 所示。

（a）

发票编号	发票日期	到期日	实际收款日期	客户ID	销售渠道	发票金额	客户名称
100002493	2016-11-04	2016-12-06	2017-01-18	1	线下经销商	43777.8	石太化工股份有限公司
100000723	2016-03-31	2016-04-29	9999-12-31	1	线下经销商	32562	石太化工股份有限公司
100000427	2016-02-23	2016-03-31	2016-03-15	1	线下经销商	6834	石太化工股份有限公司
100006812	2018-04-20	2018-06-15	2018-05-04	1	线下经销商	6834	石太化工股份有限公司
100004017	2017-05-11	2017-07-10	2017-07-03	1	线下经销商	132660	石太化工股份有限公司
100002577	2016-11-12	2016-12-28	9999-12-31	2	线下经销商	75616.2	石然天气股份有限公司

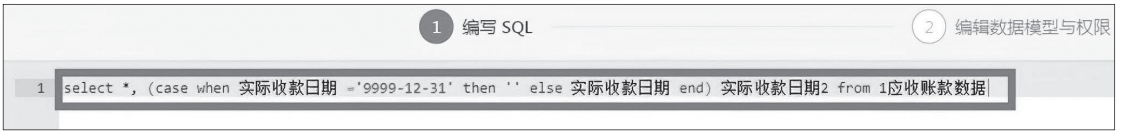
（b）

图 2-9

（8）新增列“实际收款日期 2”，将步骤（7）已完成连接的应收账款数据表与客户表合并，把该表中“实际收款日期”为“9999-12-31”的值变为空值，输入下列语句：

select *,（case when 实际收款日期 =‘9999-12-31’ then ‘ ’ else 实际收款日期 end）实际收款日期 2 from 1 应收账款数据

结果如图 2-10 所示。



（a）

销售渠道	发票金额	实际收款日期2
线下经销商	60702	2017-06-06
线下经销商	19296	2017-06-27
线下经销商	66330	
线下经销商	128318.4	2017-05-06
线下经销商	212014.8	2017-05-03
线下经销商	103273.8	

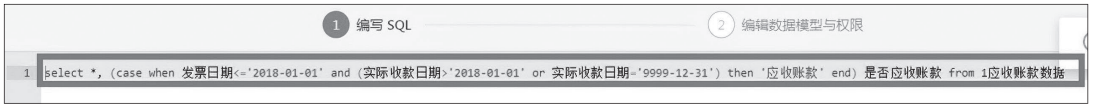
（b）

图 2-10

（9）新增列“是否应收账款”，判断每笔业务是否为应收账款。若发票日期 < =2018-01-01 和实际收款日期 > 2018-01-01 或者实际收款日期为空值，该笔业务可确定为应收账款，则“是否应收账款”的定义值为“应收账款”，其余情况为空值，输入下列语句：

select *,(case when 发票日期<='2018-01-01' and (实际收款日期>'2018-01-01' on 实际收款日期='9999-12-31') then '应收账款' end)是否应收账款 from 1 应收账款数据

结果如图 2-11 所示。



（a）

销售渠道	发票金额	是否应收账款
线下经销商	60702	
线下经销商	19296	
线下经销商	66330	应收账款
线下经销商	128318.4	
线下经销商	212014.8	
线下经销商	103273.8	

（b）

图 2-11

（10）新增列“欠款天数”并计算，设置条件为“到期日 > 2018-01-01，实际收款日期 < = 到期日”，值均为 0，到期日为 9999-12-31，值为 2018-01-01 与到期日之间的时间，其余均为实际收款日期与到期日之间的时间，输入下列语句：

```
select *,
(case when 到期日>'2018-01-01' then 0
when 实际收款日期<=到期日 then 0
when 实际收款日期='9999-12-31' then datediff('2018-01-01',到期日 )
else datediff(实际收款日期,到期日) end )欠款天数
from 1应收账款数据
```

结果如图 2-12 所示。

（a）

销售渠道	发票金额	欠款天数
线下经销商	60702	23
线下经销商	19296	27
线下经销商	66330	248
线下经销商	128318.4	0
线下经销商	212014.8	0
线下经销商	103273.8	0

（b）

图 2-12

（11）在步骤（10）的基础上新增列“账龄”，设置条件为“欠款天数 =0，值为未欠款；欠款天数 < =30，值为 1—30 天；欠款天数 < =60，值为 31—60 天；欠款天数 < =90，值为 61—90 天；欠款天数 > 90，值为大于 90 天”，如图 2-13 所示。

（a）

是否应收账款	欠款天数	账龄
	23	1-30天
	27	1-30天
应收账款	248	大于90天
	0	未欠款
	0	未欠款
应收账款	0	未欠款
	0	未欠款

（b）

图 2-13

（12）在步骤（11）的基础上新增列“预计坏账率”，设置条件为“欠款天数 = 0，坏账率为 0.05；欠款天数 < =30 天，坏账率为 0.1；欠款天数 < =60 天，坏账率为 0.2；欠款天数 < =90 天，坏账率为 0.3；欠款天数 > 90 天，坏账率为 0.5”，如图 2-14 所示。

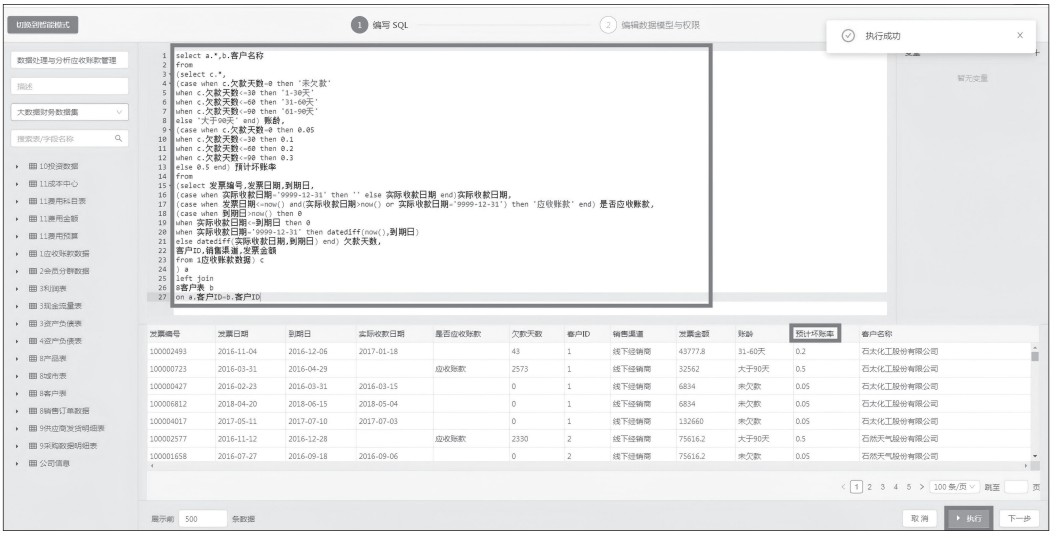


图 2-14

（13）将新表格命名为“应收账款账龄分析”，并单击“下一步”按钮，如图 2-15 所示。

（a） （b）

图 2-15

（14）对应收账款账龄分析表设置各列“数据类型”，如图 2-16 所示。

图 2-16

（15）完成对“应收账款账龄分析”表的设置，如图 2-17 所示。后续可以进行可视化分析。

图 2-17

职业要点知识测试

素质课堂

一、单项选择题

1.（　　）是指在现有数据上进行各种算法的核算，然后起到预测的作用。

A. 收集　　B. 剖析　　C. 发掘　　D. 导入

2. 大数据的处理流程不包括（　　）。

A. 数据收集　　B. 数据存储　　C. 数据处理与剖析　　D. 数据业务统计

3. 下列各项中，不属于云计算的服务模式的是（　　）。

A. IaaS　　B. DaaS　　C. SaaS　　D. PaaS

4. 离散型数据也称为（　　），其取值只能用自然数或者整数表达。

A. 连续型数据　　B. 结构化数据　　C. 不连续型数据　　D. 定序数据

5. 人的受教育程度属于（　　）。

A. 定序数据　　B. 定类数据　　C. 内部数据　　D. 外部数据

6. 下列各项中，属于大数据存储的特点的是（　　）。

A. 结构化　　B. 集中化　　C. 已稳定的数据模型　　D. 分布式

7. 下列各项中，属于非关系型数据库的特点的是（　　）。

A. 动态结构　　B. 只具备纵向扩展能力

C. 按照结构化的方法存储数据　　D. 采用表格的储存方式

二、多项选择题

1. 下列各项中，属于云计算运用的是（　　）。

A. 储存云　　B. 医疗云　　C. 金融云　　D. 教育云

2. 下列各项中，属于云计算的服务模式的是（　　）。

A. IaaS　　B. PaaS　　C. SaaS　　D. DaaS

三、简答题

1. 简述云计算的基础设施即服务。

2. 传统数据与大数据有哪些差异？

职业关键技能实训

一、实训要求

用 SQL 进行数据处理，实训原文件有两个表格，分别是应收账款数据表（见表 2-20）和客户表（见表 2-21）。

表 2-20　应收账款数据表 2

发票编号	发票日期	到期日	实际收款日期	客户 ID	销售渠道	发票金额

注：表中实际收款日期为 9999-12-31，代表暂未收款。

表 2-21　客户表 2

客户 ID	客户名称

二、实训任务

实训任务如表 2-22 所示。

表 2-22　实训任务 1

任务序号	任务内容
1	通过导入数据，建立应收账款数据表与客户表之间的关联
2	在任务 1 的基础上，新增列“实际收款日期 2”，清洗连接后的应收账款数据表和客户表的收款日期，将“实际收款日期”为“9999-12-31”的值变为空值

续表

任务序号	任务内容
3	在任务 2 的基础上新增列“是否应收账款”，判断每笔业务是否为应收账款。若发票日期 < =2019-01-01 和实际收款日期 > 2019-01-01 或者实际收款日期为空值，该笔业务可确定为应收账款，则“是否应收账款”的定义值为“应收账款”，其余情况为空值
4	在任务 3 的基础上新增列“欠款天数”并计算，设置条件为“到期日 > 2019-01-01”，实际收款日期 < = 到期日，值均为 0，到期日为 9999-12-31，值为 2019-01-01 与到期日之间的时间，其余均为实际收款日期与到期日之间的时间
5	在任务 4 的基础上新增列“账龄”，设置条件为“欠款天数 =0，值为未欠款；欠款天数 < =60，值为 1—60 天；欠款天数 < =120，值为 61-120 天；欠款天数 > 120，值为大于 120 天”
6	在任务 5 的基础上新增列“预计坏账率”，设置条件为“欠款天数 =0，坏账率为 0.05；欠款天数 < =60，坏账率为 0.1；欠款天数 < =120，坏账率为 0.3；欠款天数 > 120 天，坏账率为 0.5”