# 绪论

本章主要描述了数据分析和存储的背景，意义。分析了相关课题的研究现状，提出了本文要研究的重点内容和目标。

## 数据分析和存储的背景及意义

随着数据库技术不断更新换代，计算机信息数据不断变化，数据的重要程度与日俱增，对数据分析和存储的要求也越来越高。在当前社会中，上至企业，学校，下至个人都在进行信息化发展。而使用数据库技术可以条理分明的管理数据，使人们更加方便和迅速的管理它们，可以保持数据信息的一致性，完整性，低冗余性，甚至可以增加数据的智能性。在这样的背景下，解决对各种类型的源数据的整合提取，将数据用合适的结构存储记录的问题有着重要的意义。

## 数据领域现状及存在的问题

信息领域在近百年内飞速发展，近几年的数据又发生了巨大变化。这些变化体现在数据类型越来越多，数据结构越来越复杂，数据量越来越大。最开始的数据类型只有字符串和数字，现在又出现了图形，多媒体，cube，时间序列，流数据等类型。同时，相比以前的结构化数据，现在半结构或无结构数据越来越多，数据结构越来越复杂，对复杂数据的建模难度大大提高。而且随着数字化技术的发展，我们目前获得的数据以TB甚至PB量级迅速增长，这也是目前很火的“大数据概念”兴起的原因。然而，这样海量的数据使得原有的架构不能完全支撑，存储，处理功能都受到很严重的影响。

数据领域在过去的几十年取得了巨大的进步，然而它现在的变化也使我们面临了很多挑战。如何管理多种多样的数据类型，如何实现对复杂对象的完美建模，如何解决海量数据存储，管理，使用都是我们刻不容缓要解决的问题。

### 数据分析存在的问题和难点

数据管理系统不断发展和变化，日常生活中人们用以存储和操作数据的工具各有不同，开发平台多样化，应用系统复杂性高，再加上其他技术性问题，数据共享是常见议题，即实际项目中异构数据的存在现象十分广泛。而无论是数据挖掘，还是人工智能建模，再或者应用系统使用，在工作流里面，第一步都是接入数据，如果在第一步接入数据中，数据结构乱七八糟，那么数据的处理将遇到很多麻烦。因此我们对这些碎片化的数据整合势在必行，我们需要先进行数据分析（解析）。国外对异构数据转换研究开始的比国内早，他们目前已经在异构数据库集成，数据转换规则等上面有了较成熟的理论及技术。国内也已经开始了一些成功产品以方便数据的共享和转换，我在研究过程中有简单借鉴他们的思路。

### 数据存储存在的问题和难点

数据存储技术发展以来，一直在面临和解决着多种多样的挑战。数据显示，2010年全球存储的信息总量超过1.2ZB，2011年则继续增长了0.6ZB，而且这种增长还在加速，根据IDC2011年发布的宇宙数据研究，全球信息总量每过两年，就会增长一倍。加速增长数据总量将到达了让人心惊的程度，按照目前的速度下去，2025年将达到35ZB。大量的数据在不断产生，而他们的复杂程度，类型多样性同样也来也大，越来越多。如何实现对数据的有效存储，高效访问，方便共享和安全控制等都是信息时代非常重要的问题。

## 本文研究内容及目标

### 研究背景

本文我们将假设一个食物领域的场景，场景中有若干具体的示例，如：有品牌A,B,C生产了口味a,b,c,d等n种方便面，每种方便面有自己表层物品属性，如编号，生产日期，规格，风味等。厂商为了更好的分析食品口味，研究市场对口味的喜好，请了m名测试员从色泽，光滑性，复水性，韧性，黏性，耐泡性等多个角度进行打分。同时，在口味方面，厂商进行了细化研究，通过多种评价标准，如强度等级，麻度等级，斯科维尔指数，描述词，根据方便面在不同时间的表现进行评级，得出了味道的时间强度评价和时间频率记录。当食物投放市场后，厂商进行了反馈调查，拿到了q位用户针对n种方便面的喜恶数据。我们需要根据如上场景为厂商提供一套或多套解决方案思路，方便它们在如上场景中方便的存储和管理数据，同时保证数据处理的高效率和安全性。此为研究背景。

### 研究内容和目标

我们最终可以为厂商提供一个系统，简称为食物领域系统，可以方便它们上传文件，存储数据。为了使用方便，这个系统应该还包含简单的数据可视化界面以进行数据操作。本文研究目标是一：如何存储厂商提供的原始数据（主要是excel,txt或图片格式），并将其转换为方便进行下一步处理的统一格式。二：如何为原始数据选择合适的数据结构和数据库管理工具进行存储和管理（如何存储复杂对象；当场景不断变化时，应该如何灵活适应。）。三：为厂商提供一些简易，实用的工具，以方便非IT人士管理数据。如数据可视化，数据查找，数据管理等。

## 本文组织结构

以下的论文我们将主要探讨前两个目标问题。第二章对应数据分析问题，提出了我的一些Excel转XML想法。第三章对应数据存储问题，通过对市场上常见的几种数据库思想及其典型数据库工具的对比分析，得出不同的建模思路，对该场景下不同的示例进行了建模分析，最后还通过性能分析工具XX进行了增删改查等数据操作的性能对比。第四章是整个系统的架构，界面，功能展示，同时给出了针对第三个目标的答案。第五章是总结篇，提出了我的结论，同时抛出了本人力不能逮，尚未解决的问题。