数据库架构和关系文档

山东大学威海校区数据科学与人工智能实验班 陈其轩 王启帆 何金原

本文档用作山东大学威海校区数据科学与人工智能班陈其轩、王启帆、何金原小组的数据库与前端开发作业第 2.1 部分的阶段性数据库架构与关系文档之记录文档，包含数据库架构与关系文档两部分，各处详细位置可见目录。

目录

关于数据库架构

所谓数据库架构，就是使用什么类型的数据库，如何设计数据库，如何构建其物理模型与概念模型。DBMS 体系结构有助于数据库的设计，开发，实现和维护，数据库可以存储企业的关键信息，选择正确的数据库体系结构有助于快速安全地访问数据。

数据库设计建模步骤：概念模型->逻辑模型->物理模型

概念模型

概念模型就是在了解了用户的需求，用户的业务领域工作情况以后，经过分析和总结，提炼出来的用以描述用户业务需求的一些概念的东西。

在本次作业中，**需求为分析中国城乡差异和如何乡村振兴，用中国家庭追踪调查所有年份的数据，通过代码 从数据库调取详尽数据和用 eCharts 制作精美图表的方式，回答以下问题：中国城乡差异表现在哪里（可以细化到东中西部地区城乡对比，南北和东中西 部划分请按相关标准） --中国农村存在什么问题(可以细化到东中西部地区)。**

逻辑模型

逻辑模型是将概念模型转化为具体的数据模型的过程 ，即按照概念结构设计阶段建立的基本 E-R 图 ，按选定的管理系统软件支持的数据模型 (层次/网状/关系/面向对象)，转换成相应的逻辑模型 ，这种转换要符合关系数据模型的原则 ；

在本次作业中，我们选用关系型数据并将其转为相应的逻辑模型，具体内容将在数据库类型选取：关系型数据库与关系文档部分详细讲解。

物理模型

物理模型就是针对上述逻辑模型所说的内容 ，在具体的物理介质上实现出来，包括建什么表，建几张表。具体实现将在关系文档部分详细说明。

数据库类型选取：关系型数据库

在本次作业中，我们采用关系型数据库。

在关系型数据库系统中，关系模型包括一组关系模式，并且关系之间不是完全孤立的。如何设计一个适合的关系型数据库，其关键是设计关系型数据库的模式，具体包括数据库中应该包含多少关系模式、每一个关系模式应该包括哪些属性以及如何将这些相互关联的关系模式组建成一个完整的关系型数据库等。上述工作决定了整个数据库系统的运行效率，也是数据库系统成败的关键。

关系型数据库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，数据即表中一行一行的记录（Record），关系型数据库这一系列的行和列被称为表，表之间通过关联关系相互关联，一组表组成了数据库。用户通过查询来检索数据库中的数据，而查询是一个用于限定数据库中某些区域的执行代码。关系模型可以简单理解为二维表格模型，而一个关系型数据库就是由二维表及其之间的关系组成的一个数据组织 SQL 是关系型数据库的统一查询接口。

关系型数据库的特点

存储方式上：传统的关系型数据库采用表格的储存方式，数据以行和列的方式进行存储，要读取和查询都十分方便。

存储结构上：关系型数据库按照结构化的方法存储数据，每个数据表都必须对各个字段定义好（也就是先定义好表的结构），再根据表的结构存入数据，这样做的好处就是由于数据的形式和内容在存入数据之前就已经定义好了，所以整个数据表的可靠性和稳定性都比较高，但带来的问题就是一旦存入数据后，如果需要修改数据表的结构就会十分困难。

存储规范上：关系型数据库为了避免重复、规范化数据以及充分利用好存储空间，把数据按照最小关系表的形式进行存储，这样数据管理的就可以变得很清晰、一目了然，当然这主要是一张数据表的情况。如果是多张表情况就不一样了，由于数据涉及到多张数据表，数据表之间存在着复杂的关系，随着数据表数量的增加，数据管理会越来越复杂。

扩展方式上：由于关系型数据库将数据存储在数据表中，数据操作的瓶颈出现在多张数据表的操作中，而且数据表越多这个问题越严重，如果要缓解这个问题，只能提高处理能力，也就是选择速度更快性能更高的计算机，这样的方法虽然可以一定的拓展空间，但这样的拓展空间一定有非常有限的，也就是关系型数据库只具备纵向扩展能力。

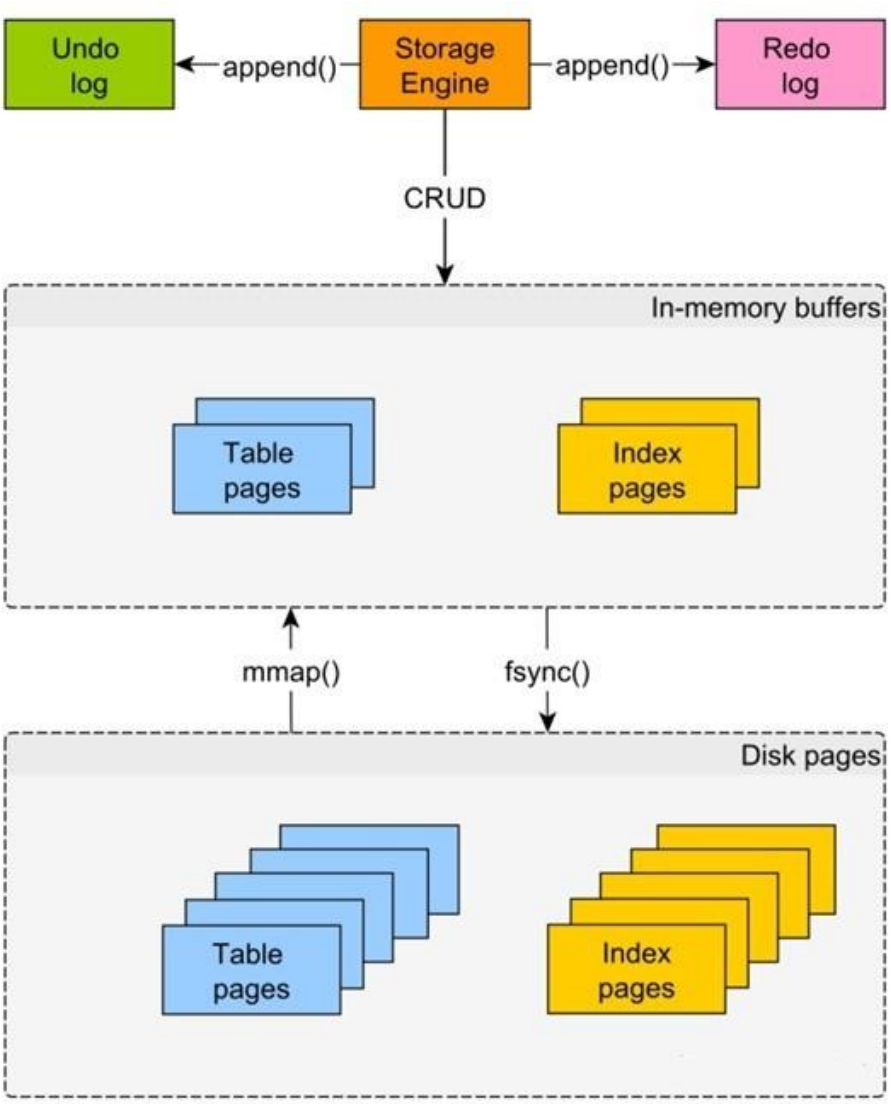
查询方式上：关系型数据库采用结构化查询语言（即 SQL）来对数据库进行查询，SQL 早已获得了各个数据库厂商的支持，成为数据库行业的标准，它能够支持数据库的 CRUD（增加，查询，更新，删除）操作，具有非常强大的功能，SQL 可以采用类似索引的方法来加快查询操作。

规范化：在数据库的设计开发过程中开发人员通常会面对同时需要对一个或者多个数据实体（包括数组、列表和嵌套数据）进行操作，这样在关系型数据库中，一个数据实体一般首先要分割成多个部分，然后再对分割的部分进行规范化，规范化以后再分别存入到多张关系型数据表中，这是一个复杂的过程。好消息是随着软件技术的发展，相当多的软件开发平台都提供一些简单的解决方法，例如，可以利用 ORM 层（也就是对象关系映射）来将数据库中对象模型映射到基于 SQL 的关系型数据库中去以及进行不同类型系统的数据之间的转换。

事务性：关系型数据库强调 ACID 规则（原子性（Atomicity）、一致性（Consistency）、隔离性（Isolation）、持久性（Durability）），可以满足对事务性要求较高或者需要进行复杂数据查询的数据操作，而且可以充分满足数据库操作的高性能和操作稳定性的要求。

特别的，关系型数据库的架构设计，主要是要解决存储和事务。存储是要解决数据的查询问题。而事务则包含了四个特性：原子性、一致性、隔离性、持久性。

下图就是一个关系型数据库的典型架构。其中索引的存在是为了提高数据查询的性能：



关系型数据库选取：MySQL

在本次作业中，我们选用 MySQL 数据库（关系型数据库管理系统）并将其布置于服务器上（Server version: 8.0.27 MySQL Community Server - GPL）。其中部署于云服务器上的 MySQL 部署于腾讯云标准 CVM 上。

MySQL 是一个关系型数据库管理系统，由瑞典 MySQL AB 公司开发，目前属于 Oracle 公司。MySQL 是一种关联数据库管理系统，关联数据库将数据保存在不同的表中，而不是将所有数据放在一个大仓库内，这样就增加了速度并提高了灵活性。

- MySQL 是开源的，目前隶属于 Oracle 旗下产品。
- MySQL 支持大型的数据库。可以处理拥有上千万条记录的大型数据库。
- MySQL 使用标准的 SQL 数据语言形式。
- MySQL 可以运行于多个系统上，并且支持多种语言。这些编程语言包括 C、C++、Python、Java、Perl、PHP、Eiffel、Ruby 和 Tcl 等。

- MySQL 对 PHP 有很好的支持，PHP 是目前最流行的 Web 开发语言。
- MySQL 支持大型数据库，支持 5000 万条记录的数据仓库，32 位系统表文件最大可支持 4GB，64 位系统支持最大的表文件为 8TB。
- MySQL 是可以定制的，采用了 GPL 协议，你可以修改源码来开发自己的 MySQL 系统。

关系模式的范式：BC 范式

在本次作业中，我们将采用 BC 范式范式（高于一般企业开发需求）。在中国家庭追踪调查（China Family Panel Studies，CFPS），也即本次的数据集中，主码，候选码，函数依赖等均是明确的，例如问卷调查隐去了所有个人信息，事实上每个问卷的编号就是唯一的主码。例如成人问卷与儿童问卷，其主码便是第一个属性 pid(person id)。而在家庭问卷中，类似于此的编码为 fid，同样的，村居问卷就是 cid。

BC 范式（BCNF）是 Boyce-Codd 范式的缩写，其定义是：在关系模式中每一个决定因素都包含候选键，也就是说，只要属性或属性组 A 能够决定任何一个属性 B，则 A 的子集中必须有候选键。. BCNF 范式排除了任何属性 (不光是非主属性，2NF 和 3NF 所限制的都是非主属性)对候选键的传递依赖与部分依赖。

规范的叙述，即是：**设关系模式 R<U，F>∈1NF，如果对于 R 的每个函数依赖 X→Y，若 Y 不属于 X，则 X 必含有候选码，那么 R∈BCNF。**

函数依赖：直接函数依赖

在本次作业中，对于任何调查问卷主码皆唯一，**例如成人问卷与儿童问卷，其主码便是第一个属性 pid(person id)。而在家庭问卷中，类似于此的编码为 fid，同样的，村居问卷就是 cid。**也即是其中的每一函数依赖 X→Y 全为直接函数依赖

BC 范式分解算法

（1） 把所有的属性构成一个关系模式
（2） 如果关系模式不是 BCNF，找到里面的函数依赖，X->A，如果 X 不是键，那么删除 A，并将 XA 划为新的子集（此时该子集的依赖左侧都是键）
（3） 都是 BCNF，否则继续分解剩余元素集。

求属性闭包的算法

设 R<U，F>，A 为 U 中属性(集)。
(1) X(0)=X

(2) X(i+1)=X(i)∪A

(3) 其中：对 F 中任一个 Y->A ，且 Y⊆X(i)；求得 X(i+1) 后，对 Y->A 做删除标记。若 X(i+1)=X(i) 或 X(i+1) =U 则结束，否则转(2)。

关系文档

CFPS 概述

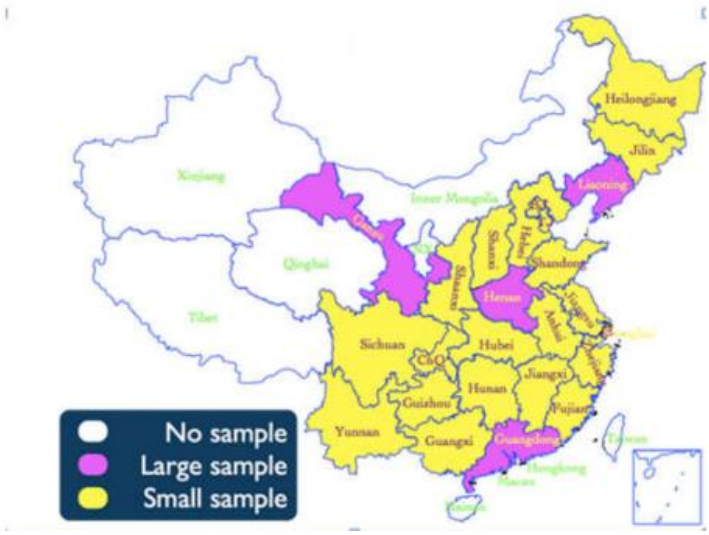
- **中国家庭追踪调查(China Family Panel Studies,CFPS)**
- **研究内容：中国社会的变化　　(Panel)**
- **研究载体：家户（Family）**
- **研究主题：经济、教育、健康、幸福…**

- 中国家庭追踪调查（CFPS）是一项全国性、综合性的社会追踪调查项目，旨在通过追踪收集个体、家庭、社区三个层次的数据，反映中国社会、经济、人口、教育和健康的变迁，为学术研究和公共政策分析提供数据基础。
- CFPS 于 2007 年开始始前期工作，2008、2009 年开始进行预调查。2010 年开始正式调查，此后每 2 年进行一次常规的跟踪调查。(除了在 2011 年还对部分样本进行了一次小规模的样本维护调查）因此共有 2010、2011、2012、2014、2016、2018 年五年的数据，目前还没有公布 2020 年的。(年份之间问卷的差异不大，下面的例子包括官方给的用户手册的例子都是基于 2010 年的)
- 数据的来源（抽样）：将在后续部分详谈。
- 关于问卷的设计：将在后续部分详谈。
- 问卷中碰到的属性名，都可以在对于年份的问卷或是用户手册的第七节中查询到。
- 用户手册的第八章对部分数据(包括家庭收入、城乡分布、受教育程度、婚姻状态……)做了初步的分析和评估，对于重要数据的选择和分析具有较大的参考价值。

抽样过程

- CFPS 最初目标样本规模为 16000 户，其中，有 8000 户从上海、辽宁、河南、甘肃、广东五个独立子样本（称为“大省”）过度抽样（oversampling）得到，每个“大省”1600 户。另有 8000 户则从其他 20 个省份共同构成的一个独立子样本框（称为“小省”）抽取。
- 五个“大省”的子样本具有地区自代表性，可以进行省级推断以及地区间比较。5 个“大省”样本框在二次抽样后，与“小省”样本框共同构成具有全国代表性的总样本框。(就是说因为大省抽的多故不能直接来推断全国，因此和小省做比较时候要再进行一次二次抽样）(也即如果是推断全国的话，属性 subsample 那里的值应该是 1）

- 五个大省从地理上正好包含了五个方位。



五类问卷

- CFPS 的主体问卷包括村居问卷、家庭成员问卷、家庭问卷、少儿问卷和成人问卷五类。
- 村居问卷
- 家庭成员问卷
- 家庭（经济）问卷
- 少儿问卷
- 成人问卷

  注：村居问卷属于社区一层，家庭成员问卷和家庭（经济）问卷属于家庭一层，少儿问卷、成人问卷都属于个人问卷一层（三层是由大到小的）。

村居问卷

- 仅 2010 年有，之后的年份都没有。
- 主要的受访对象是比较了解村居、能接触到统计资料的人员，由这些人员尽量多地回答这份问卷。
- 村居问卷的主要目的是了解村（农村社区）或者居（城市社区）的设施、人口、政治、经济、历史、政策等相关情况。共有 ABCDEFGHJKYZ 共 12 个模块的内容。
- 关于表中数据和题目的对应说明：如表中的属性名 ca1 表示村居问卷 A 模块第一题，ca3_s_2 表示 A 模块第三题（多选题）的第二个空，cb3 表示 B 模块第三题。

模块	2010 年问卷内容	2014 年问卷内容
A 基础设施	村/居属性，受访人职务，设施，地界，行政面积，水源，燃料，高污染企业	村/居属性，受访人职务，设施保有及新增，地界，行政面积，水源，燃料，高污染企业
B 人口结构	总户数，总人口，户籍人口，常住人口，外来流动人口，年龄结构，出生与死亡，少数民族	总户数，总人口，户籍人口，常住人口，外来流动人口，年龄结构，出生与死亡，少数民族，大姓分布
C 社会保障与食品价格	低保政策，物价水平	低保政策，物价水平
D 行政	村居办公人员规模，办公条件，周边交通，选举情况	村居办公人员规模，选举情况
E 历史政治	历史变革，是否为旅游区，是否有高污染企业，最近一次村居委会选举情况	
F 房屋价格	商品房历史最高价、上个月最高价、上个月一般价	商品房历史最高价、上个月最高价、上个月一般价
G 环境、交通与资源	距最近集镇、县城、省城的距离与交通时间，矿产资源，自然灾害，土地资源	距最近集镇、县城、省城的距离与交通时间，矿产资源，自然灾害，土地资源
H 劳动力、产值与收入	劳动力结构，雇工价格，农业总产值，非农业总产值，人均纯收入，大姓分布	劳动力结构，雇工价格，农业总产值，非农业总产值，人均纯收入
J 医疗与生育	医疗点面积，医疗卫生人员数量，农村合作医疗开展情况，计划生育政策	医疗点面积，医疗卫生人员数量，农村合作医疗开展情况，二胎政策
K 财政状况	集体企业及产值，财政总收入及来源，财政总支出及支出项目	征地经历、集体财政总收入及来源，财政总支出及支出项目，债务情况
Y 受访者资料	受访者性别、年龄、政治面貌与受教育程度，村/居主任性别、年龄、政治面貌与受教育程	受访者性别、年龄、政治面貌与受教育程度，村/居主任性别、年龄、政治面貌

家庭成员问卷

- 家庭成员问卷的回答人必须满足两个基本条件：一是同灶吃饭的成员之一；二是与家庭具有血缘/婚姻/领养关系。
- 家庭成员问卷的主要目的是界定样本家庭的内部关系网络。

家庭（经济）问卷

- 2010、2012、2014 年叫家庭问卷，2016 年起改名为家庭经济问卷。
- 家庭问卷的主要目的是在家庭层面上收集样本家庭的日常生活、社会交往与经济活动方面的信息。共有 ABCDEFGHJKZ 共 11 个模块的内容。

模块	问卷内容
A 地理交通	最近的公交、医疗点、高中、商业中心
B 生活条件	用水，燃料，电，卫生间条件，垃圾处理，保姆/小时工雇佣
C 社会交往	春节拜访，送礼，族谱/家谱，祭祖/扫墓，邻里交往，亲友交往
D 住房情况	房屋所有权，自建/购买，租房来源，建筑面积，入住时间，房屋市值与租金，房屋结构，其他房产情况，住房困难情况
E 经营状况[26]	U 外出工作模块（外出人员，工作地址，时间投入，假期是否回家，转移支付情况，家庭是否因其外出而雇佣/增加帮工），政府补助，致贫原因，V 非农经营模块（非农产业类型、数量、参与者、总资产、家人拥有股份、雇佣人数，营业额，税后纯利润），房屋出租，土地与其他生产资料出租，财物出卖，拆迁，土地征用
F 家庭收入	存款，金融产品，离退休金/社会保障金/低保收入，工资/奖金/补贴/红利等收入，非工资性/农业生产收入，礼金/礼品折现
G 家庭资产	保险可赔偿额，他人欠款，收藏品价值，其他资产现值
H 家庭支出	最贵消费品花费，借贷款，家庭各项日常支出（食品、出行、通信等），家庭各项特殊支出（家电、医疗保健、教育、商业保险等），捐赠，总支出
J 耐用品	汽车，摩托车，拖拉机，电视
K 农业生产	土地类型，土地数量，农业收支状况，农林作物类型、产量、销量、收入，家畜与渔业类型、产量、销量、收入，家畜饲养条件
Z 访员观察	问卷回答人，家庭住房条件，家庭整洁度，家庭成员精神面貌，家庭成员间关系，长幼关系，性别间关系，受访者个人特征

少儿问卷

CFPS 将 16 岁以下的人群定义为少儿，16 岁及以上的人群定义为成人。少儿问卷与成人问卷是分别为这两类人群设计的个人访问问卷。

- **教育：**
  不在上学人群——教育史、不上学原因、毕业阶段、所学专业、今后打算、教育期望
  正在上学人群——**上学模块**（***共用***）：所处阶段、学校类型、专业、学习成绩、课外辅导、学生活动、学习与学校情况主观评价、教育期望
- **工作经历：**是否有过正式工作、工作内容、工作报酬、时间投入
- **时间利用**（***共用***）：生活、工作、学习培训、娱乐休闲与社会交往、交通活动
- **人际交往与日常生活：**交友情况、K歌跳舞等娱乐活动、恋爱关系、家务劳动、零花钱
- **手机与网络**（***共用***）：生活、工作、学习培训、娱乐休闲与社会交往、交通活动
- **身体健康：**健康、饮食、锻炼
- **个人经历：**是否坐过火车/飞机、对时政的了解
- **主观评判：**自尊量表、成就量表、价值量表、抑郁量表、亲子关系，等等
- **认知测试：**方向、识字、数学
- **其它：**社会压力、职业期望、幸福感，等等

成人问卷

	住的时间	
兄弟姐妹情况	兄弟姐妹数目、名字、出生日期、是否健在、去世年龄与原因、婚姻状况、最高学历、职业、行政/管理职务、居住地，在世父母和谁住在一起，去世父母的去世原因	（仅在 CFPS 2010 使用）
教育史	已完成的最高学历，小学至已完成的最高学历之间的各个阶段的学校类型、学习时间、结束时间、学校名称、是否毕业、学科与专业等，教育期望	基线题组
语言运用	各类语言重要程度，与家人交流的语言	核心题组
上学模块	当前正在上学所处阶段，学校类型，专业，学习成绩，课外辅导，学生活动，学习与学校情况主观评价，教育期望	核心题组
婚姻	婚姻状况（未婚/在婚/同居/离婚/丧偶），现任/前任/初婚配偶/同居对象的出生年月、结婚/同居时间、婚前同居情况、如何认识，前次/初次婚姻解体的原因与时间	核心题组
子女关系	60 岁以上受访者与子女关系评价，与子女间的交往活动	核心题组
工作	见图 12	核心题组
个人收入	非经营性收入，经营收入，亲友资助，国家政府补贴救济	核心题组
时间利用	生活，工作，学习培训，娱乐休闲与社会交往，交通活动	核心题组
娱乐休闲	闲暇活动，频率，出行方式，出国经历	核心题组
手机网络	手机使用情况，社交网络使用、邮箱使用情况，网络重要性评估，上网频率与地点	核心题组
社会关系	找人帮忙，烦恼倾诉，社会地位自评	核心题组
主观测量	价值观，社会观，成就量表，生活满意度等	核心题组
政治	遭遇偷抢威胁的经历，不公正待遇，新闻关注，政府工作评价	核心题组
健康	身高，体重，健康自评，身体不适，慢性疾病，住院经历，医疗费用，病痛处理方式，对医疗状况的满意度，中医，体育锻炼，饮食，P-ADL，吸烟喝酒经历，睡眠，记忆力，生病时主要照料人，身体机能	核心题组
心理健康	K6 量表、CESD 量表	轮替题组