医疗商保理赔场景下的数据汇聚实验

# 实验目的

1. 掌握数据汇聚的基本流程
2. 了解且学会使用数据汇聚中涉及到的技术
3. 培养在实际场景中灵活使用所学知识的能力
4. 能够借助互联网资源，自主学习和运用相关知识

# 场景描述

目前，医疗健康的商业保险的理赔环节，需要保险用户自己提供的材料非常复杂。除了要提供申请人的身份证以外，还需要 提供能支持索赔的全部账单、证明、信息和证据，比如病历资料、诊断书、处方、病理检查报告、化验检查报告、出院小 结、医疗费用原始单据、费用明细单据等。同时整个过程需要人工上传相关的记录和证明材料，并且由保险公司进行线下核 实，整个理赔流程环节繁琐、审核周期长，用户体验十分不友好。

为打通商保服务民生最后一公里，某市卫健委、市数据局、数据资产运营有限公司、保险公司等多方参与，共同探索开展线 上快保核赔业务，该业务经个人授权同意后，利用卫健委已汇聚的医疗机构的诊疗记录、发生费用、电子病历等医疗数据信 息，通过数据要素在不同主体之间的流通，实现在线理赔核保、理赔服务，同时探索开展普惠保险产品设计，为老百姓提供 更多优质普惠的医保产品，实现“数据多跑路，群众少跑腿，商保更普惠”。

要完成以上任务，首先我们需要将不同来源的数据进行汇聚并生成数据目录，以方便对各项数据资产的管理。

# 数据说明

本项目数据集来自于四个主体：

1. ，包括门诊数据、住院数据、检查数据、费用数据等



医院

1. ，主要为医院基本信息、疾病描述信息

卫健委

1. ，包括居民基本信息、医保参保信息、医保理赔结算数据

人社局

1. ，包括商保产品数据、商保参保信息、商保理赔结算数据

保险公司

数据保存在

data

文件夹中，更详细的数据信息介绍请参考

readme

文件。需要指出，本次实验使用的数据均为合成数

据，意味着其中具体的实体之间的对应关系可能并不符合实际。

# 实验要求

1.

基本要求

2.

拓展要求

# 实验步骤

：按照步骤完成数据汇聚的基本任务，最终生成实验报告文件，并附上必要的代码与运行测试结果

：按照步骤提示，完成拓展任务

## 任务一：医院主体数据的标准化与对齐(数据对齐)

不同医院的数据表的格式有部分差别，为了更好使用相关数据，卫健委指定了标准的数据模式，现要求你将A、B医院的数 据与标准数据模式进行对齐，并生成新的数据表(csv文件)

标准的数据模式的格式如下： 门诊数据表 ( )

OutpatientData

|  |  |  |
| --- | --- | --- |
| **属性名** | **数据类型** | **示例数据** |
| 医院名 | TEXT | 医院A |
| 门诊ID | VARCHAR | OP12345 |
| 病患姓名 | VARCHAR | 张晓明 |
| 病患身份证号 | VARCHAR | 430221 19930809 8118 |
| 就诊日期 | DATE | 2023-06-15 |
| 就诊时间 | TIME | 20:33 |
| 就诊科室 | VARCHAR | 内科 |
| 诊断结果 | VARCHAR | 上呼吸道感染 |
| 医生ID | VARCHAR | D001 |
| 医生姓名 | VARCHAR | 王萧 |
| 处方信息 | TEXT | 阿莫西林 500mg\*3天 |
| 诊疗费用 | DECIMAL | 150.00 |

住院数据表 ( )

InpatientData

|  |  |  |
| --- | --- | --- |
| **属性名** | **数据类型** | **示例数据** |
| 医院名 | TEXT | 医院A |
| 住院ID | VARCHAR | IP12345 |
| 病患姓名 | VARCHAR | 张晓明 |
| 病患身份证号 | VARCHAR | 430221 19930809 8118 |
| 入院日期 | DATE | 2023-06-10 |
| 出院日期 | DATE | 2023-06-20 |
| 住院时间 | INT | 10 |
| 诊断结果 | VARCHAR | 肺炎 |
| 负责科室 | VARCHAR | 内科 |

|  |  |  |
| --- | --- | --- |
| **属性名** | **数据类型** | **示例数据** |
| 医生ID | VARCHAR | D002 |
| 医生姓名 | VARCHAR | 王萧 |
| 住院费用 | DECIMAL | 5000.00 |

检查数据表 ( )

ExaminationData

|  |  |  |
| --- | --- | --- |
| **属性名** | **数据类型** | **示例数据** |
| 医院名 | TEXT | 医院A |
| 检查ID | VARCHAR | EX12345 |
| 病患姓名 | VARCHAR | 张晓明 |
| 病患身份证号 | VARCHAR | 430221 19930809 8118 |
| 日期 | DATE | 2023-06-12 |
| 时间 | TIME | 20:33 |
| 类型 | VARCHAR | 血液检查 |
| 费用 | DECIMAL | 50.00 |
| 医生ID | VARCHAR | D002 |
| 医生姓名 | VARCHAR | 王萧 |

费用数据表 ( )

ExpenseData

|  |  |  |
| --- | --- | --- |
| **属性名** | **数据类型** | **示例数据** |
| 医院名 | TEXT | 医院A |
| 费用ID | VARCHAR | EXP12345 |
| 患者姓名 | VARCHAR | 张晓明 |
| 患者身份证号 | VARCHAR | 430221 19930809 8118 |
| 缴费日期 | DATE | 2023-06-15 |
| 缴费时间 | TIME | 20:33 |
| 费用类型 | VARCHAR | 药品费 |

|  |  |  |
| --- | --- | --- |
| **属性名** | **数据类型** | **示例数据** |
| 金额 | DECIMAL | 100.00 |

请你使用数据对齐技术，给出你的匹配规则以及最终生成的新的数据表。此外，后续任务中，请将用新生成数据表替代医院 A与医院B的数据表。

tips ：自动匹配的方法是很有可能出错的，故你在自动生成匹配方式后，还需自己进行比对，并改正错误内容(请你具体在

报告中说明哪些为你使用的自动匹配技术能识别正确的，哪些为识别错误的)。你也可以尝试多种自动匹配的算法以找到最 好的。

## 任务二：医疗商保场景下数据模型的建立(数据对象化)

为了清晰地确定医疗商保场景下的数据实体和相关信息，将复杂且离散的数据集合转化为具体、清晰、可操作的数据对象， 请按照实际场景描述，将数据进行对象化，生成ER模型图

本任务一般由业内专业人士与数据库管理员人工完成，但目前出现了结合大语言模型辅助进行数据图谱建设的方法。因此， 本任务的 为自行进行相关资料查询，尝试使用大语言模型自动进行进行数据对象化。

拓展任务

## 任务三：医疗商保数据的标签体系建立(数据标签化)

为了方便商保主体对于病人信息的访问，我们需要将属于病人这一数据对象的相关数据进行标签化。具体要做的任务如下：

首先，我们需要对病人的相关数据字段(即属性)进行聚类，如“姓名”、“性别”可以划分为基础信息，“住院时长”、 “所患疾病”可划分为诊疗信息。

然后，根据商保主体进行保险理赔时的标签需求，建立需要的统计标签与算法标签，具体需要给出标签名与标签计算方法。 最后，请将标签信息进行整理，生成一个标签文档。

商保进行理赔时需要判断的部分内容如下： - 病人所患疾病的疾病分级 - 病人的参保信息 - 病人进行诊疗的总金额 - 病人的财产收入水平(判断是否符合保险公司特殊关照条例。为了简化问题，可以月收入10000为界限) - 病人是否为政府保障人员(判断是否符合保险公司特殊关照条例)

：自行进行相关资料查询，尝试使用大语言模型自动进行进行数据标签化。

拓展任务

## 任务四：医疗商保数据的分类分级(数据分类分级)

请按照实际情况，将数据进行分类分级。最终结果可按照下表的形式提供

|  |  |  |  |
| --- | --- | --- | --- |
| **数据类别** | **类别描述** | **具体数据项** | **风险级别** |
|  |  | [该类别具体包含的数据] | [核心、重要或一般数据] |

## 任务五：医疗商保数据目录(数据目录建立)

根据以上数据汇聚各项任务的分析结果，提供全面、易于查找、便于理解和值得信赖的数据资产清单。