大致要求:

◉ 同一lab两人组队，每个小组独立完成，一个小组只需要一个人交report。第一次和第二次proj队友是同一个

◉ 在ddl前交

◉ 不要抄袭

◉ 报告内容要求8-20页

我们的第一次proj主要分为如下四个方面

1. 基于提供的数据关系和数据文件设计E-R图
2. 根据提供的数据文件用PostgreSQL设计一个关系型数据库
3. 把**所有的**数据导入数据库
4. 比较**数据库**和**原始文件的IO**在**数据检索**和**操纵**的性能。编程语言允许C/C++，Java，Python

背景

这次的数据库用来存储**市场部门的组织结构**和销售合同的**细节。**
老师提供了**市场部门的组织结构**和需要被导入进**数据库中的数据样例**其中数据样例包含员工和order的信息

顾客来自世界上的七个区域

Europe, America, Asia, Eastern China, Northern China, Southern China, and Southwestern China

每个区域都由一个供给中心掌管,每个供给中心的负责人在如下图中
注意：①Asia这个供给中心不会处理China的所有事务
②hongkong，macau，taiwan算作中国的境外事务，即归Asia管

Europe	America	Asia	Eastern China	Northern China	Southern China	Southwestern China
Audrey Evans	Miriam Evans	Steven Edwards	Xu Zhuyu	Kong Yibo	Yang Penglong	Tao Yibo

在所提供的数据集合中，有从5_000个合同中得到的5_0000条销售记录（即一个合同会包含多条销售记录）

数据描述：

data提供了5_000个合同，每个合同有**1或多个**order，(Each order represents a single kind of product that has been ordered) (**这个原文没看懂**)每个product都被分解成多个order，总共有5_0000个order，下面是column的具体解释

1. contract number: 每个合同的unique标识符; 每个的值大概如右边所示CSEXXXXXXX, 从 CSE0000000 到 CSE0004999.
2. client enterprise: client enterprise的名字
3. supply center: 与client enterprise相对应的供给中心
4. country: client enterprise所在的国家
5. city: client enterprise所在的城市

6. industry: client enterprise所在的行业
7. product code: 该product的unique标识符；每个的值是有数字和字母组成的六位数字符串，e.g., L8N0649, C186H47, M40V792.
   注意：每个product code有且仅有一个product name，有超过一个的product_model
8. product name: product的名字
9. product model: product的具体型号。**注意：**每个project model有其自己的unit price（我的理解是每个product有多个unit price，视型号而定)
10. unit price: 本合同中该product的单价
11. quantity: 在合同中要求的product的数量
12. contract date: 创建合同的日期
13. estimated delivery date: 预期交付product的时间
14. lodgement date: 实际的product交付日
15. director: 负责这个合同的负责人
16. salesman: 做这个order的salesman的名字
17. salesman number: salesman的number（号码）
18. gender: salesman的性别
19. age: salesman的年龄
20. mobile phone: salesman的电话号码

注意：

◉ 如果client enterprise不在China，则city的值为null

◉ 如果lodgement date的日期在2022-3-2之后，则其值是null

report的要求

1. 名字，学号，lab班级
2. 百分比，写清楚

Task 1 E-R Diagram(15' pt)

◉ 用**任意画图软件**画出我们设计的数据库的E-R图。绝对不允许手写，一定按照E-R图的标准格式画图

◉ 在report中提供E-R图的截图即可，同时标注我们用来画图的software/online service

Task 2 Database Design(25' pt)

基于上面提供的信息设计tables 和 columns，通过datagrip的“Show Visualization” 功能生成E-R图。简要解释tables和columns的意义，在report中，我需要提供如下内容

◉ 用datagrip自带的功能生成的E-R图的截图

◉ 简要解释tables和columns的意义

同时，以附件的形式上交包含DDL（create table 语句）的文件，要和report分开交

注意：

1. 所有的数据是基于`contract_info.csv`这个文件得到的
2. 我们的设计需要遵守前三个范式

3. 用primary key来指代重要的属性，用foreign key来指代关系
4. 在每个table中的每一行都需要unique，并且能被primary key标识 (可用simple key也可用composite key)
5. 每个表要有外 键，或者有其他表的外键指向
6. 对于表之间的外键方向，不能有环。 例如：A表有外键关联B表，B表有外健关联C表，C表有外健关联A表
7. 每个表至少有一个“Not NULL”的column（包括primary key 但不包括id column）（**没懂**）
8. 除了主键自增的id之外，需要有其他unique约束的列
9. 对每个属性要用合适的数据类型
10. 当需求改变的时候，我们的设计需要易于拓展

Task 3 Data Import(25' pt)

通过脚本语言把`contract_info.csv`这个文件里面的内容导入我们之前设计的数据库中。在我完成"导入"操作之后,我需要保证我导入成功,在这个task中,我需要完成以下基本需求(15' pt in 25' pt):

◉ 完成写入数据的脚本

◉ A description of how you use the script to import data. You should clearly state the steps, necessary prerequisites, and cautions in order to run the script and import data correctly
描述我是如何利用脚本导入数据的,我需要极为清晰地按顺序 阐述必要的步骤和注意事项

还需要完成如下附加内容(10' pt in 25' pt)

◉ 找到超过一种方式导入数据,并且提供数据分析,a comparative analysis of the computational efficiencies(不同方法之间的效率?)

(不懂)

◉ 尝试优化我们写的脚本,描述一下我们优化后相比于优化前带来的速度提升

对于附加内容,保证提供我们每次的`测试环境` `程序` `实际的时间花销`,我们需要写一或者两个段落来分析实验结果(可以参照task 4的具体细节)

Task 4 Compare DBMS with File I/O (35' pt)

比较database APIs和file APIs在同一个编程语言条件下,数据导入和操作的性能差别,注意要按照如下步骤来比较分析:

1. 使用database APIs进行基准测试(benchmarking):
   首先,准备一个有至少2_0000行的表(我最好复用本节课用到的表,或者这次proj用到的表),然后用编程语言通过database APIs(包括 Insert Delete Update Select)来访问数据库.我可以指定每类语句的数量.最终我需要记录每条语句和同类型语句的运行时间
2. 使用File APIs进行基准测试(benchmarking):
   和1中一模一样,不过需要用标准的 File APIs.
   首先,准备一个文件,其内容需要和1中的数据一模一样,然后用File APIs进行操作,每个语句的数量,每种类型语句的数量和1中一样
3. 比较分析: 分别比较记录相同的`operation/statement`我可以从不同级别来比较:
   ①.我可以分析比较单个语句
   ②.我可以分析比较同一类型的语句

在report中我需要完成如下基本内容(20' pt in 35' pt)

◉ 描述我的测试环境,包括

1.硬件: 包括但不限于CPU型号,内存大小,使用的是固态硬盘还是传统硬盘

2.软件: 包括但不限于我的DBMS的版本,操作系统,我选择的编程语言,开发环境(语言的版本,编译器的版本,库的版本)

3.当某人尝试去复制我的实验时,我需要给他提供什么必要的信息

◉ 描述我是如何在DBMS和数据文件中组织我的测试数据的,包括tables的创建和文件的数据格式

◉ 描述我的测试脚本和我的源代码,不要粘贴复制源代码,不过需要用附件的形式上交

◉ 比较数据的时候,需要用可视化的形式表达出来(用数据的形式),除了把运行时间以图表的形式表示之外,我还需要描述关于运行表现的主要区别,我在结果中知道到了什么有意思的东西,我们实验中的亮点

Some notes on how to finish this task in a better way:

1. 操作的数量不应该很小(如少于100条insert,少于10条select)
2. 我可以选择我想要的数据方式来存在文件中,如纯文本形式(CSV,JSON,XML等等)或者自定义的二进制形式
3. 一定要只用标准I/O流,如java中的java.io,或者Python中的file对象,唯一的例外是当我选择使用XML或者JSON时,我可以用相关的第三方库,(Gson for Java, the json package in Python, etc.)
4. 有许多第三方库可以用于加速(Pandas in Python),但是在使用那些库之前,我们需要先对比DBMS和标准I/O流的差别
5. 一些可能有用的网站

   9 Advantages of Database Management System over File System (csestack.org)

   Advantages of Database Management System (tutorialspoint.com)

   Chapter 3 Characteristics and Benefits of a Database – Database Design – 2nd Edition (opentextbc.ca)

在report中我需要完成如下附加内容(15' pt in 35' pt)

1. High concurrency and transaction management**(高并发和啥玩意)**
2. 用户权限管理
3. Database index and file IO**(啥玩意)**
4. 更优雅的方式来展示我们实验的结果
5. 不同的数据库软件间,软件系统,编程语言,库,操作系统间的对比