PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
EDUCACIÓN PROFESIONAL - ESCUELA DE INGENIERÍA
Dipĺomado en Ciencia de Datos para Políticas Públicas

Curso: DCDPP - Gestión de Datos
Fecha: 5 de junio de 2021
Profesor: M. Arancibia
Ayudante: F. Alessandri

Tarea 1

Fecha de Entrega: Sábado 19 de junio - 23:59:00

En las clases de gestión de datos hemos aprendido acerca de la eficiencia de bases de datos, modelos relacionales, visualizar información y unir diferentes bases de datos para obtener la mayor información posible. Esta tarea consistirá de dos preguntas las cuales deberá responder de forma completa (código, texto e imágenes explicativas). Puede entregar la tarea en el formato, que estime conveniente (doc, pdf, script de R, url de Colab) pero tenga en cuenta que debe adjuntar el código utilizado. Se recomiendo el uso de Colab R.

Problema 1. (1.5 puntos)

En el taller, revisamos el siguiente código para hacer un *join* de dos bases de datos que compartieran un identificador.

```
1   ciclo1 <- 3000
2    ciclo2 <- 3000
3    n <- 1
4
5    for(i in 1:ciclo1){
6    mrun <- rendim2017$mrun[i]
7     for(j in 1:ciclo2){
8        mrun2 <- rendim2018$MRUN[j]
9          if(mrun == mrun2){
10            fila <- c(as.character(rendim2017[i,]), as.character(rendim2018[j,]))
11            rendim_17_18[n, ] <- fila
12            n <- n+1
13        }
14    }
15    }
```

Suponga que usted tiene un disco duro con dos bases de datos con 3000 filas cada una que ocupan 300 paginas de disco cada una, en otras palabras cada pagina tiene 10 tuplas. Por otro lado usted tiene una memoria RAM de 101 páginas de almacenamiento.

(a) (0.5 pts) Dibuje el diagrama de flujo del algoritmo.

(b) (0.5 pts) Cuente la cantidad de veces que el disco duro va buscar los datos a memoria de la forma que esta escrito acá el algoritmo.

(c) (0.5 pts) Plantee y programe una forma para reducir las lecturas a disco a al menos 1300 páginas. El algoritmo que requiera menos consultas del curso tendrá un bonus de 0.5 puntos.

Problema 2. (1.5 puntos)

(a) (0.7 pts) Descarga la base de datos de Cargos Docentes del año 2015 en R. Crea dos tablas distintas: una paraa el establecimiento y la otra para los docente, además agrega con los 4 atributos más importantes en

cada relación. Además crea una tabla que vincule el docente con el establecimiento. Notese que buscamos reducir el tamaño de la información por lo que en las tablas no deberían aparecer valores duplicados.

(b) (0.4 pts) Utilizando los operadores de álgebra relacional escribe la siguiente consulta para las tablas que definiste en el punto anterior:

Liste los nombres de los profesores que trabajan en más de un colegio.

(c) (0.4 pts) Implementa esta consulta en R (puedes usar el procedimiento dado por el álgebra relacional pero no es obligatorio).

Problema 3. (3 puntos)

En esta pregunta usted deberá trabajar con las bases de datos disponibles dentro del portal de datos para investigación del MINEDUC (`datosabiertos.mineduc.cl`). En esta pregunta el objetivo es comprobar alguna hipótesis basándose en visualización de información, por lo que las elecciones de cada paso deben ser concordantes.

(a) (0.5 Pts) Revise la página web del MINEDUC y formule a lo menos tres distintas hipótesis que puedan responderse con los estos datos. Estas deben cumplir que:

- ◇ Deben resolverse mediante el cruce de información entre 2 o más tablas (no es valido usar la misma base de datos para distintos años).
- ◇ No deben ser triviales pues se evaluará también la calidad de la pregunta.

(b) (0.5 Pts) Seleccione tres bases de datos que ayuden a resolver una o más hipótesis planteadas. Explique su elección y con sus propias palabras que contiene cada base de datos seleccionada.

(c) (1 Pt)Compruebe usando R que las bases de datos se pueden mezclar entre si. Es decir, pueden hacer un join entre ellas cuyo resultado es distinto de vacío. Si es que este proceso no entrega la cantidad de tuplas suficiente para estudiar ninguna de las tres hipótesis (o al menos tener confianza del resultado) deberá rehacer los pasos (a), (b) y (c).

(d) (1 Pt) Seleccione una de las tres hipótesis y a través de la visualización de información determine si esta pareciese estar acertada o no. Entregue una pequeña conclusión y análisis sobre este mini-estudio.