



# Estructura de Datos y Algoritmos

ITBA 2024-2

El código para leer un csv con header, a través de esta biblioteca podría ser así. Bajar de campus **DataAnalyis.java** y verificar que se obtiene el dump correcto.

```
public class DataAnalysis {  
    public static void main(String[] args) throws IOException {  
  
        // leemos el archivo  
        // URL resource = DataAnalysis.class.getClassLoader().getResource("co_1980_alabama.csv");  
        URL resource= DataAnalysis.class.getResource("/co_1980_alabama.csv");  
  
        Reader in = new FileReader(resource.getFile());  
        Iterable<CSVRecord> records = CSVFormat.DEFAULT  
            .withFirstRecordAsHeader()  
            .parse(in);  
  
        // imprimimos los registros con todos sus valores  
        for (CSVRecord record : records) {  
            String value = record.get("daily_max_8_hour_co_concentration");  
            System.out.println(String.format("%s, %s", value, record.toString()));  
  
        }  
        in.close();  
    }  
}
```



Qué sucede si colocamos otro iterador a continuación?

La variable “records” es una colección accesible?

Si lo que queremos realizar frecuentemente es:

- Buscar el promedio de la polución registrada.
- Imprimir ascendentemente la info disponible, pero ordenada por polución.
- Averiguar si existió una polución cuyo valor fuera 2.8
- Buscar el valor numérico de la mínima polución registrada.
- Buscar la info disponible en que se dio la mínima polución registrada
- Conocer qué valores numéricos de polución se registraron entre [3.65, 3.84]
- Conocer la info disponible en la que la polución registrada fue entre [3.65, 3.84)
- Conocer la info disponible en la que la polución registrada fue [10.5, 12]

En cuál/cuáles de esas operaciones conviene hacerlo por medio del Index y no analizando los datos tal cual vienen?

Sobre cuál de los campos definiríamos el orden del índice?

RTA:

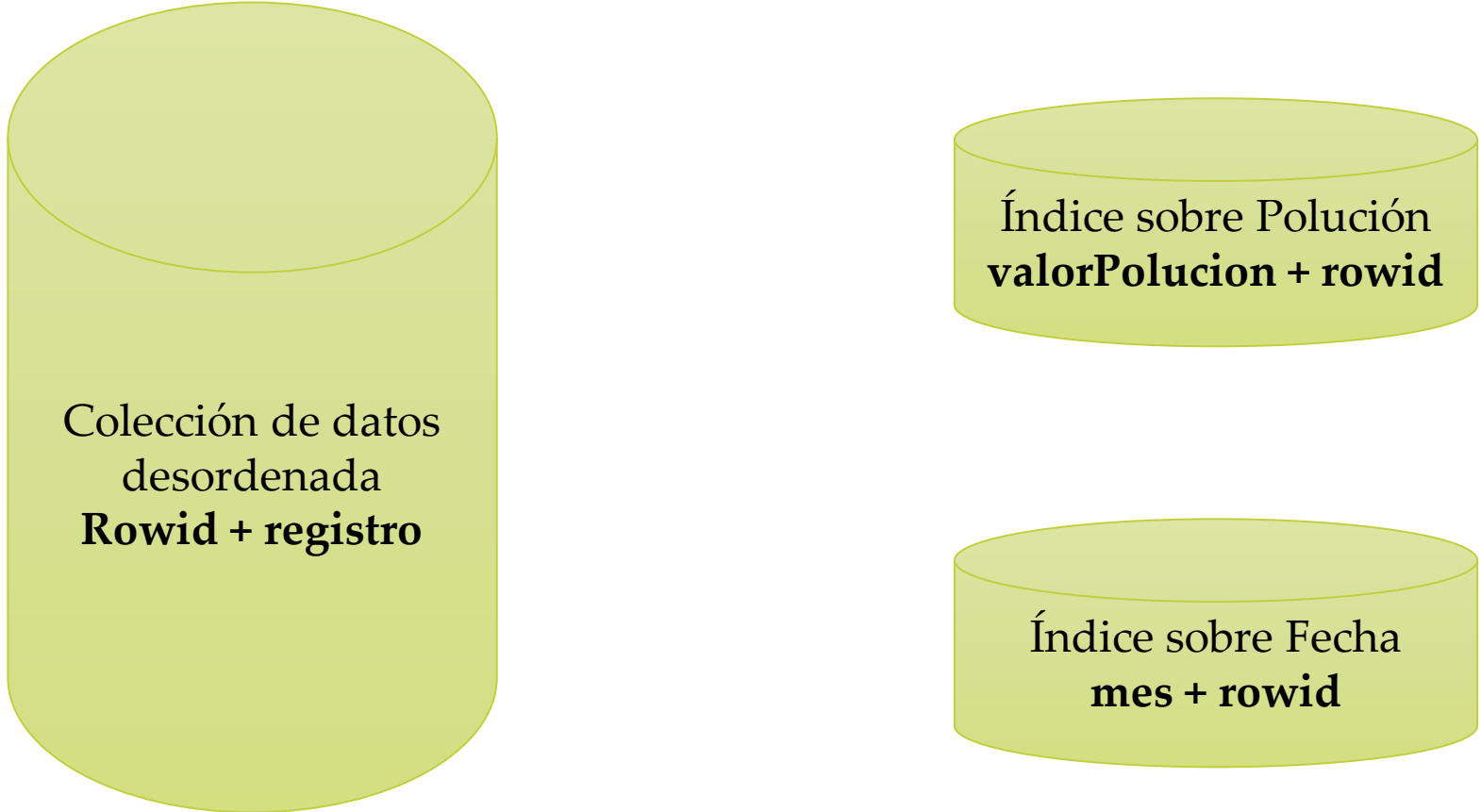
Sobre el campo: "daily\_max\_8\_hour\_co\_concentration"

## Como funcionan los índices en colecciones de datos?

### Consideraciones

- La **colección de datos** (en este caso la línea del csv) tienen muchísimas componentes, pero sólo sobre unas pocas de ellas se quiere realizar búsquedas por rango, etc.
- Algunas operaciones son favorecidas por búsqueda binaria (imprimir ordenado, rango, etc). La **colección de datos** no puede ordenarse simultáneamente por diferentes campos. O está ordenada por polución o está ordenada por fecha, etc. Como la colección tiene muchísimas componentes “duplicar”, “triplicar”, la información es costoso.

- Pero si creamos una estructura auxiliar con la mínima información necesaria para llegar a ella? Esa información **tiene que favorecer las operaciones** que esperamos y puede permitir resolver sólo con ella algunas consultas o llevarnos sólo hacia los elementos esperados (los que satisfacen la consulta) en la colección de datos grande  $\Rightarrow$  INDICE
- Un índice favorece las operaciones solo sobre la CLAVE DE BUSQUEDA.



The diagram illustrates a database structure with three cylinders. The largest cylinder on the left represents the main data collection. To its right are two smaller cylinders, one above the other, representing indexes. The top right cylinder is an index on 'Polución' (Pollution), and the bottom right cylinder is an index on 'Fecha' (Date). All cylinders are light green with a darker green outline.

Colección de datos  
desordenada  
**Rowid + registro**

Índice sobre Polución  
**valorPolucion + rowid**

Índice sobre Fecha  
**mes + rowid**

# Caso de uso

## Colección de datos desordenada

1 100, Oct 2020,.....  
2 350, May 2020, .....  
.....  
55555 2, May 2020,.....  
.....  
99999 3, Feb 2021,.....


## Índice sobre Polución

2 55555  
3 99999  
100 1  
350 2

## Índice sobre Mes

Feb 99999  
May 2  
May 55555  
Oct 1





La colección de datos, según lo que estamos pretendiendo hacer con ella, ¿En qué tipo de colección Java la podríamos guardar?

Rta

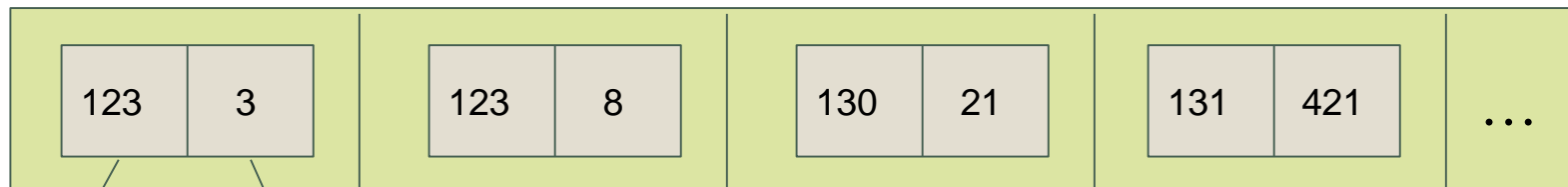




Para implementar esos requerimientos, modificaremos  
DataAnalysis:

- Bajar de campus la clase **IdxRecord.java** que nos permite representar un Registro<T1, T2> donde T1 es el key de ordenamiento. Lo demás es la info que queremos asociar. Pueden agregarle métodos.
- Resolver cada una de los requerimientos.

Cada ítem del índice va a guardar dos valores:  
la key por la que ordenamos y un valor adicional  
asociado.



Key

Value

Preparar los datos para las consultas

```
// coleccion de valores
```

```
HashMap<Long, CSVRecord> datos= new HashMap<>();
```

```
// indice sobre polucion o los que deseemos
```

```
IndexParametricService<IdxRecord<Double, Long>> indexPolucion=  
new IndexWithDuplicates<>(IdxRecord.class);
```

```
// coleccion de datos
```

```
for (CSVRecord record : records) {
```

```
    // insertamos en la colección y en indice
```

```
    COMPLETAR!!!!
```

```
}
```

**CONSULTAS!!!!**

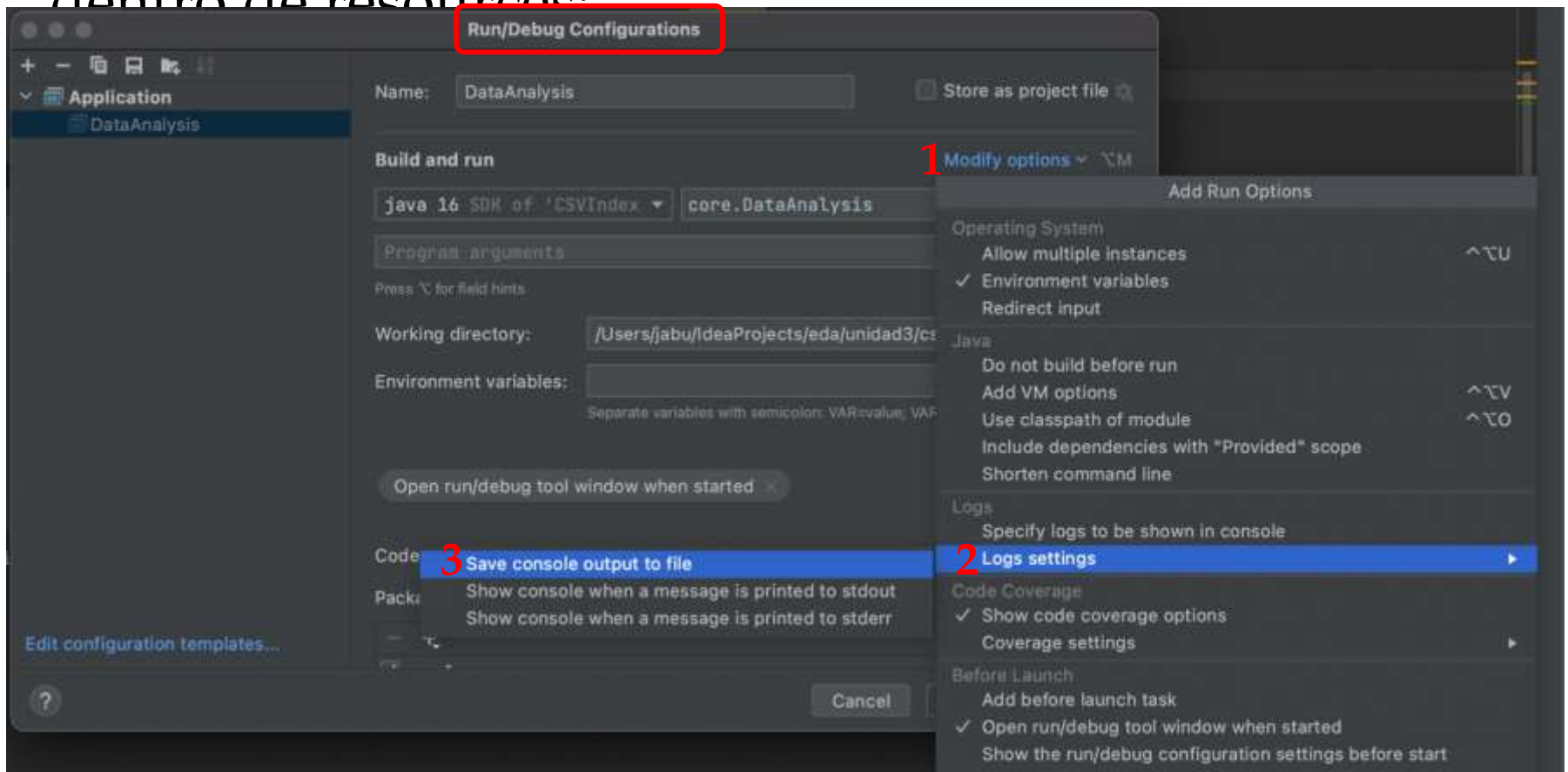
# Detalles de implementación

**CSVRecord** provee el método **getRecordNumber()**

Recordar que en el **HashMap** van los datos completos asociados a un id  
mientras que en el índice va el valor por el que quiero buscar asociado al id.

Recordar insertar en el **Hashmap** y en el **índice**

Si los datos son muchos, puede que la consola nos muestre solo los últimos valores. Si quieren, redireccionen la salida, por ejemplo en un archivo dentro de resources:




- Caso de Uso: Averiguar si existió una polución cuyo valor fuera 2.8

...

Se puede usar solo el índice?

Preciso también de datos?

...



Debería dar True


- Caso de Uso: Buscar el valor numérico de la mínima polución registrada

...

Se puede usar solo el índice?

Preciso también de datos?

...



Debería dar 0.3

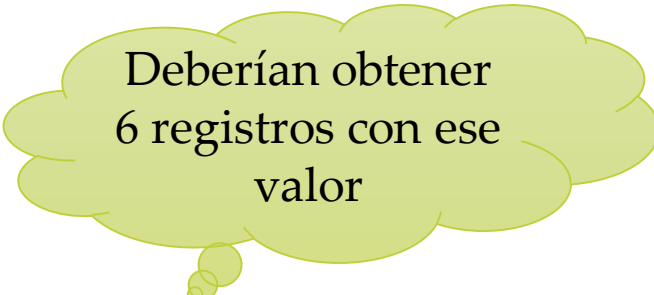


- Caso de Uso: Buscar el valor numérico de la mínima polución registrada e imprimir de ese mínimo TODA la información asociada

...

Se puede usar solo el índice?

Preciso también de datos?



Deberían obtener  
6 registros con ese  
valor

...

```
promedio registrado: 2.3268163804491406
CSVRecord [comment='null', recordNumber=407, values=[1980-07-20, 10731003, true, 0.3, ppm, 3, 24,
CSVRecord [comment='null', recordNumber=646, values=[1980-07-27, 10890014, true, 0.3, ppm, 3, 24,
CSVRecord [comment='null', recordNumber=165, values=[1980-09-21, 10730027, true, 0.3, ppm, 3, 24,
CSVRecord [comment='null', recordNumber=421, values=[1980-08-03, 10731003, true, 0.3, ppm, 3, 24,
CSVRecord [comment='null', recordNumber=724, values=[1980-11-27, 10890014, true, 0.3, ppm, 3, 24,
CSVRecord [comment='null', recordNumber=164, values=[1980-09-20, 10730027, true, 0.3, ppm, 3, 24,
```

- Caso de Uso: Imprimir los valores de polución ordenados ascendentemente.

Se puede usar solo el índice?

Preciso también de datos?

....

Deberían aparecer

los valores

numéricos

almacenados

desde 0.3 hasta el

9.5

- Caso de Uso: Imprimir TODA la info pero ascendentemente ordenada por polución

Se puede usar solo el índice?

Preciso también de datos?

....

Deberían aparecer  
los registros  
ordenados por los  
valores de  
polución  
desde 0.3 hasta el  
9.5

- Caso de Uso: Conocer qué valores numéricos de polución se registraron entre [3.65, 3.84]

...

Se puede usar solo el índice?

Preciso también de datos?

...

Deberían obtener

3.7

3.7

3.8

3.8

3.8



Si las consultas que queremos realizar frecuentemente son:

- Si en la Latitud 34.68776113 hay sensores instalados;
- Imprimir los valores de latitudes donde hay sensores instaladas, ordenadas ascendentemente;
- Imprimir la info donde los valores de las latitudes es mínima;
- etc.

En qué campo habría que crear el index?

