Pontificia Universidad Católica de Chile
Departamento de Ciencia de la Computación
IIC2413 - Bases de Datos

Examen

Fecha: 9 de julio de 2025

1º semestre 2025 - Profesores: Eduardo Bustos - Christian Álvarez

---

INSTRUCCIONES

- La prueba tiene 3 preguntas más un bonus, cada una con diferente puntaje. El puntaje total de la prueba es 60 puntos. La nota del examen es $(\sum_{i=1}^{bonus} P_i)/10 + 1$.
- **Lea toda la prueba primero**, las preguntas no están en orden de dificultad.
- Cada hoja de respuesta debe contener su **nombre completo** y **número de lista** en la parte superior.
- Responda cada pregunta en una hoja diferente.
- Debe firmar la lista de asistencia.
- Duración: 2.5 horas.
- Al finalizar la prueba debe **digitalizarla y subirla a Canvas** correctamente. No seguir esta instrucción los expone a un descuento de hasta 5 décimas.

Pregunta 1 - Modelo ER, Diseño de BD, SQL (30 pts)

Usted trabaja para una empresa de evaluación de impacto ambiental y le encargaron la misión de recolectar datos de un conjunto de estaciones de monitoreo, armar una base de datos con ellos, calibrarlos y generar unas consultas.

Esta misión tiene varios desafíos que cumplir, entre las que están la obtención de los datos que están publicados en las páginas de las instituciones, la conversión de los datos al mismo sistema de medidas (unidades métricas), entre otros. A continuación se describe la situación con la que se encuentra.

En la Figura 1 se muestra una estación de monitoreo y su ubicación en el terreno (cuadrantes).

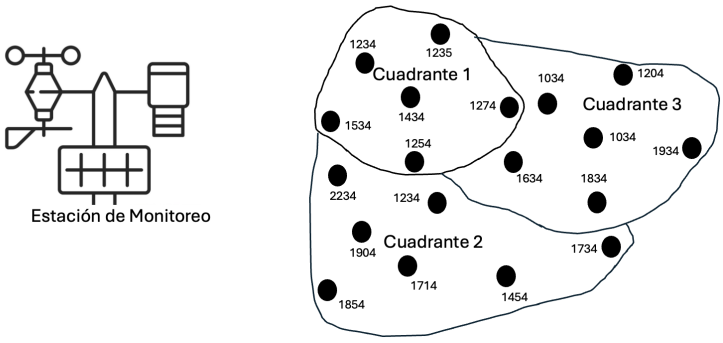

Figura 1: Estación de monitoreo y ejemplo de cuadrantes

Cada estación registra una variedad de **datos** (correspondientes a **parámetros** como temperatura, humedad del aire, presión atmosférica, etc.), cuya cantidad varía entre estaciones. Por ejemplo, una estación puede medir 2 parámetros ambientales y la otra 5 o más. Sin embargo, en los archivos de datos vienen siempre todos los parámetros y en el mismo orden. La frecuencia de medición de datos en todas las estaciones es el mismo, 5 minutos.

Para acceder a los datos, se deben descargar los archivos CSV desde las instituciones que reportan las estaciones, teniendo en cuenta que la misma estación puede ser reportada por más de una institución, y además cada institución puede usar diferentes unidades de medida para los parámetros. Cada archivo CSV reporta **una sola estación.** A continuación se describen los nombres de los archivos, parámetros, formatos y unidades:

- Sobre las estaciones: La lista de las estaciones de monitoreo está disponible en el sitio web de las instituciones, y tiene el nombre `listadeestaciones.csv`. Estos contienen, para cada estación, el ID único, nombre, dueño, ubicación única en coordenadas y tecnología. La tecnología e institución definen el sistema de unidades (métrico o Imperial). Por ejemplo: *12323, FISUR, SAGO, -40.57397, -73.13509, THX1138*.
- Sobre los archivos CSV de mediciones: Como se mencionó, todos los archivos de

mediciones contienen todos de parámetros existentes en el mismo orden, independiente de la tecnología de la estación. En el caso de que una estación no provea de un parámetro, este dato aparece nulo en el CSV.
Por ejemplo: *2025-04-28 12:05:10, 30, 50, 105, 200, , 3*.

Cada nombre de archivo CSV de monitoreo está estandarizado y se compone del ID de la estación, el nombre de la institución de origen y la fecha de extracción del reporte.
Por ejemplo: `12323-INIA-28042025.csv`.

- Sobre los parámetros: Los datos en el archivo CSV de mediciones y sus rangos válidos son (recuerde que las unidades dependen de la institución de origen del reporte):

Parámetro	**Descripción**	**Rango válido**	**Unidad**
fecha	Fecha y hora de la muestra	–	aaaa-mm-dd hh:mm:ss
Vviento	Velocidad del viento	[0, 200]	kn o km/h
Dviento	Dirección del viento	[0, 360]	grados
T	Temperatura	[-100, 150]	°F o °C
R	Radiación	≥ 0	W/m^2
P	Presión atmosférica	≥ 0	hPa o $mbar$
LL	Precipitación acumulada	≥ 0	oz o ml

Nota: Los datos se miden en forma absoluta en el momento de la muestra, excepto Precipitación acumulada, donde el valor se incrementa en cada medición (o sea, el nuevo valor es mayor o igual al anterior), y se reinicia a las 00:00 de cada día.

I. Modelo ER y Esquema relacional

a) (3 pts.) Cree un **modelo entidad relación** con todos los elementos vistos en clase (entidades, relaciones, cardinalidad, claves primarias, parciales y foráneas, entidades débiles, etc.).

Solución: Solo es necesario 3 entidades, una débil, 2 relaciones. Si hay más entidades deben ser útiles al problema.

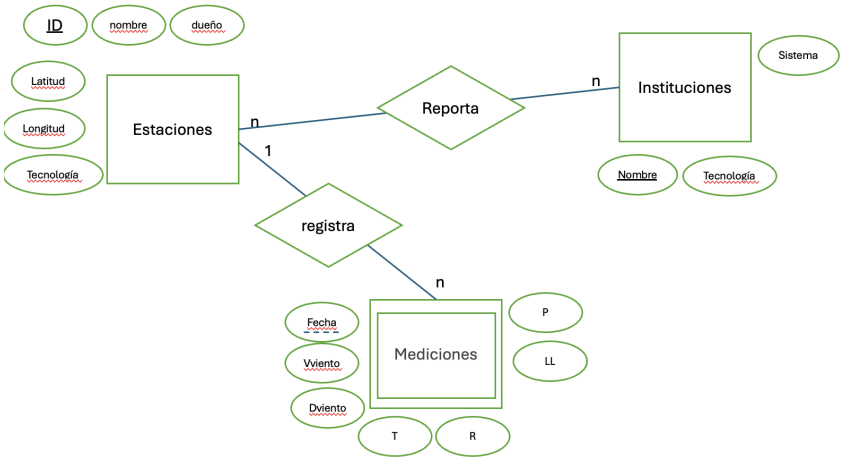

Figura 2: Diagrama E/R.

b) (5 pts.) Cree el **esquema relacional** en 3NF para el modelo ER. Incluya los nombres de las tablas, atributos, dominios, llaves y restriciones de integridad. Recuerde prevenir la duplicación de datos (una estación puede estar incluida en más de un reporte).

Solución: Puede escribir el esquema como DDL o según lo visto en clase. Lo importante son los atributos, tipos, dominios y sobre todo las restricciones de integridad de los dominios, PK y FK.

Estaciones						Institución				
	ID	INT	PK			Nombre	TEXT	PK		
	nombre	TEXT				Tecnología	TEXT	PK	FK Estaciones(tecnología)	
	dueño	TEXT				Sistema	ENUM ('IMPE	'METRICO')		
	Latitud	FLOAT								
	Longitud	FLOAT								
	Tecnología	TEXT								
Mediciones										
	ID	INT	PK	FK references Estaciones(ID)						
	fecha	DATETIME/TIMESTAMP	PK							
	Vviento	INT CHECK (Vviento BETWEEN 0 AND 200)								
	Dviento	INT CHECK (Dviento BETWEEN 0 AND 360)								
	T	INT CHECK (T >= -100 AND T <= 150)								
	R	FLOAT <Check(R>=0)								
	P	FLOAT <Check(P>=0)								
	LL	FLOAT <Check(LL>=0)								

Figura 3: Esquema relacional.

II. Carga de datos en la instancia

c) (4 pts.) Como indica el enunciado, alguna información viene en archivos CSV, pero otra está publicada en páginas. Indique el **procedimiento** para transformar el problema

del enunciado en datos de tablas del esquema mencionando el origen de la información, tablas de destino y método de carga (manual, automatizada). Justifique cada acción.

Solución:

- Desde el sitio web de las instituciones (una o más da igual) se descarga la lista de estaciones.
- Desde el sitio de las instituciones se descarga el archivo de las medidas.
- Desde el sitio web se extrae manualmente la información para armar la table de unidades de medida (Tecnología, Institución, Medida).

  Nota: Como el ID de la estación está en el **nombre** del archivo y no en los datos, será necesario un proceso de 2 etapas para cargar la información (está más adelante).

d) (3 pts.) Cree los **SP** necesarios para transformar los datos que tengan unidades distintas a las métricas a sus equivalentes de este sistema (°F → °C, oz → ml, hPa → mbar, nudos → km/h) según corresponda e indique la sentencia para insertarlo en la BD. Considere:

$$
\begin{aligned}
C &= \frac{5}{9} \cdot (F - 32) \\
\text{ml} &= \text{oz} \cdot 29{,}5735 \\
\text{mbar} &= \text{hPa} \\
\text{km/h} &= \text{nudos} \cdot 1{,}852
\end{aligned}
$$

Solución: Están separados las 3 conversiones cada uno en un SP, y luego otro que los agrupa. Puede hacerse todo en uno solo. Sean blandos con la sintaxis, evalúen como si fuera pseudocódigo.

```
CREATE OR REPLACE PROCEDURE fahrenheit_a_celsius(
IN valor_fahrenheit NUMERIC,
OUT valor_celsius NUMERIC
)
LANGUAGE plpgsql
AS $$
BEGIN
valor_celsius := 5/9 * (valor_fahrenheit - 32);
END;
$$;


CREATE OR REPLACE PROCEDURE onza_a_ml(
IN valor_onza NUMERIC,
```

```
OUT valor_ml NUMERIC
)
LANGUAGE plpgsql
AS $$
BEGIN
valor_ml := valor_onza * 29.5735;
END;
$$;


CREATE OR REPLACE PROCEDURE nudo_a_kmph(
IN valor_onza NUMERIC,
OUT valor_ml NUMERIC
)
LANGUAGE plpgsql
AS $$
BEGIN
valor_ml := valor_onza * 1.852;
END;
$$;


CREATE OR REPLACE FUNCTION cambio_unidades()
RETURNS TRIGGER AS $$
BEGIN
NEW.T := fahrenheit_a_celsius(NEW.T);
NEW.P := onza_a_ml(NEW.P);
NEW.V := nudo_a_kmph(NEW.V);

RETURN NEW;
END;
$$ LANGUAGE plpgsql;
```

e) (3 pts.) Cree los **Triggers** necesarios para que al insertar datos en (°F, oz, hPa, nudos) ejecute el SP de cambio de unidad correspondiente (°C, ml, mbar, km/h).

Solución: La conversión debe hacerse solo para los archivos con sistema de unidades Imperial. Lo importante es que el trigger debe estar BEFORE insert.

```
CREATE TRIGGER convertir_unidades_mediciones
BEFORE INSERT ON mediciones
FOR EACH ROW
EXECUTE FUNCTION cambio_unidades();
```

f) (4 pts.) Cree, mediante sentencias SQL o comandos de PSQL, la **carga de los datos** desde los archivos CSV a la instancia. Los datos de monitoreo se deben cargar en unidades métricas. Para hacer esto use los SP y Triggers creados en d) y e).

Solución: Crea una tabla temporal para cargar datos y cargar cada uno de los archivos

```
\copy monitoreoTMP FROM '12323-INIA-28042025.csv' WITH (FORMAT csv)
```

Luego para cada archivo insertar los valores agregando el número de la estación.

```
INSERT INTO monitoreo (id, fecha, Vviento,Dviento, T, R, P, LL)
    SELECT 123, fecha, temperatura, humedad
FROM monitoreoTMP;

\copy estaciones FROM listadeestaciones.csv WITH (FORMAT csv)
```

III. Consultas

Aunque los parámetros entregados estén dentro de su rango, pueden estar fuera de su comportamiento usual (por ejemplo, el 9 de julio T=36°C). Por eso, se realiza una **calibración** que verifica si los valores están dentro del rango aceptable según la fecha, usando un registro histórico de los parámetros por **cuadrante**. Un cuadrante agrupa varias estaciones, y cada estación pertenece a un solo cuadrante (ver Figura 1). Para esto considere las tablas:

- Cuadrante(id, id_estacion)
- HistoricoTemperatura(fecha, id_cuadrante, PromT, sdT, PromV, sdV, PromR, sdR, PromP, sdP)

Donde fecha indica el día-mes-año, PromX es el valor promedio diario del parámetro X, sdX la desviación estándar para ese día, T=Temperatura, V=Velocidad del viento, R=Radiación, P=Presión. No es necesario calibrar la Precipitación acumulada, esta se copia desde la tabla de Mediciones.

g) (4 pts.) Cree una **consulta** que lea cada uno los parámetros (T, V, R, P) desde la tabla de mediciones y entregue en una tabla MedicionesCalibradas los datos calibrados si cumplen con la condición y un valor que represente un error en caso contrario. Un valor se considera calibrado si está en el rango $\mathbf{PromX} \pm 2 \cdot \mathbf{sdX}$, y erróneo si está fuera. Indique los valores usados para representar error para cada parámetro.

Solución:

```
INSERT INTO MedicionesCalibradas (idestacion, fecha, Vv, Fv, T, V, R, P, LL)
SELECT
m.id_estacion,
m.fecha,
((m.T BETWEEN h.PromT - 2 * h.sdT AND h.PromT + 2 * h.sdT) * m.T) +
((m.T < h.PromT - 2 * h.sdT OR m.T > h.PromT + 2 * h.sdT) * -999.9)
```

```
AS T_calibrada,
((m.V BETWEEN h.PromV - 2 * h.sdV AND h.PromV + 2 * h.sdV) * m.V) +
((m.V < h.PromV - 2 * h.sdV OR m.V > h.PromV + 2 * h.sdV) * -99.9)
AS V_calibrada,
((m.R BETWEEN h.PromR - 2 * h.sdR AND h.PromR + 2 * h.sdR) * m.R) +
((m.R < h.PromR - 2 * h.sdR OR m.R > h.PromR + 2 * h.sdR) * -9999.9)
AS R_calibrada,
((m.P BETWEEN h.PromP - 2 * h.sdP AND h.PromP + 2 * h.sdP) * m.P) +
((m.P < h.PromP - 2 * h.sdP OR m.P > h.PromP + 2 * h.sdP) * -999.9)
AS P_calibrada
FROM
Monitoreo, HistoricoTemperatura, Cuadrante
WHERE Cuadrante.idestacion = Monitoreo.id_estacion
    AND DATE(Monitoreo.fecha) = HistoricoTemperatura.fecha
    AND Cuadrante.id = HistoricoTemperatura.id_cuadrante;
```

h) (4 pts.) Cree una **consulta** que entregue los valores promedio diario calibrado para cada parámetro y estación. No es necesario tomar en cuenta los datos erróneos.

Solución:

```
SELECT
  idestacion,
  fecha,
  AVG(NULLIF(T_calibrada, -999.9)) AS Promedio_T,
  AVG(NULLIF(V_calibrada, -99.9)) AS Promedio_V,
  AVG(NULLIF(R_calibrada, -9999.9)) AS Promedio_R,
  AVG(NULLIF(P_calibrada, -999.9)) AS Promedio_P
FROM MedicionesCalibradas
GROUP BY id_estacion, fecha
ORDER BY id_estacion, fecha;
```

Pregunta 2 - Evaluación de consultas (15 pts)

Se tienen las siguientes 2 relaciones:

- Pedidos(id_pedido, fecha_pedido, monto, estado)
- Despachos(id_despacho, id_pedido, fecha_despacho)

Cada página puede almacenar 20 tuplas de **Pedidos** y 50 tuplas de **Despacho**. Además el buffer almacena 12 páginas.

a) (4 pts.) ¿Cuántas lecturas de páginas se requieren para realizar un Nested Loop Join entre **Pedidos** y **Despachos** usando esta tabla como relación externa?

Solución:

$$\text{Costo NLJ} = \text{Pág. de Pedidos} + (\text{Tuplas de Pedidos} \times \text{Pág. de Despachos})$$

$$\text{Pág. de Pedidos} = \frac{\text{Tuplas de Pedidos}}{20}$$

$$\text{Pág. de Despachos} = \frac{\text{Tuplas de Despachos}}{50}$$

b) (4 pts.) Si se utiliza Merge Join, asumiendo que las relaciones están ordenadas por **id_pedido**. ¿Cuál sería el costo?

Solución:

$$\text{Costo Merge Join} = \text{Pág. de Pedidos} + \text{Pág. de Despachos}$$

$$\text{Pág. de Pedidos} = \frac{\text{Tuplas de Pedidos}}{20}$$

$$\text{Pág. de Despachos} = \frac{\text{Tuplas de Despachos}}{50}$$

c) (4 pts.) ¿Hay diferencia si se cambia la tabla que se usa como relación externa? Justifique su respuesta.

Solución: Sí hay diferencia, ya que cambia principalmente el número de veces que se recorre la tabla externa. Por lo tanto conviene elegir como tabla externa la que tiene más tuplas, ya que ese valor se disminuye al determinar las páginas que usa la tabla.

d) (3 pts.) ¿Cambian los valores calculados anteriormente si no hay un índice creado para el campo **id_pedido**? Justifique su respuesta

Solución: Al no existir un índice, al realizar el join entre dos tablas hay que hacer un barrido completo por la tabla (full scan), lo cual incrementa la cantidad de operaciones que se deben realizar por cada valor buscado.

Luego, Costo = Tuplas Pedido $\times$ Tuplas Despacho.

Pregunta 3 - Privacidad (15 pts)

a) (4 pts.) ¿Cuáles de las siguientes condiciones se deben cumplir para que un análisis de dato preserve la privacidad? Justifique su respuesta.

   I. El análisis tiene k-anonimato mayor o igual a 50.

   II. El análisis tiene l-diversidad.

   III. No se viola la privacidad de algún individuo.

   IV. Se puede aprender algo del análisis.

   V. Usa Privacidad Diferencial.

   Solución: Al analizar cada una de las afirmaciones tenemos lo siguiente:

   - I: Aunque el k-anonimato es una medida que aumenta la privacidad, no lo asegura.
   - II: Aunque l-diversidad es una medida que aumenta la privacidad, tampoco lo asegura.
   - III: Esta es una condición que debe cumplirse, ya que asegura que no se ha violado la privacidad de ningún individuo.
   - IV: Al cumplir la privacidad, debe ser factible el aprendizaje a partir de los datos. De nada sirve tener privacidad, sin aprender de los datos.
   - V: La privacidad diferencial es un mecanismo que incrementa la privacidad, pero no es requisito, ya que se pueden usar otros mecanismos.

   Por lo tanto la respuesta correcta es III y IV.

b) (6 pts.) Responda las siguientes consultas con respecto a la privacidad diferencial.

   I. ¿Qué es la privacidad diferencial?

   Solución: La privacidad diferencial es un mecanismo que garantiza que un algoritmo aleatorio opera sobre un conjunto de datos, no revele información específica de ningún individuo.

   II. ¿Por qué la privacidad diferencial ayuda a mantener la privacidad?

   Solución: Porque asegura que la probabilidad de obtener un resultado particular es casi la misma, independientemente de si un individuo está o no en el conjunto de datos.

   III. Explique un ejemplo práctico donde aplique la privacidad diferencial.

   Solución: Por ejemplo tenemos una base de datos de 100 personas, dónde 25 están enfermos. Si consulto que me indique el número total de enfermos y después

elimino un individuo de la muestra, podré saber si esa persona está o no enferma.

Al incorporar a los datos privacidad diferencial, no me entregará el resultado exacto (porque contiene ruido) y al sacar un individuo de la muestra no estaré seguro si está enfermo o no.

c) (5 pts.) Relacione el concepto y su definición.

i. De-identificación (DeId)

ii. Datos sintéticos (DS)

iii. Re-identificación (ReId)

iv. k-Anonimato (kAn)

v. l-Diversidad (lDiv)

Solución:

	Definición	**Concepto**
1	Para cada grupo de filas con los mismos CIds, aparece al menos un cierto número de veces repetidos.	kAn
2	Para cada grupo de filas con los mismos CIds, por cada atributo sensible S, existen un cierto número de valores distintos.	lDiv
3	Es el proceso que vuelve a asociar una persona con un conjunto de datos.	ReId
4	Es el proceso que remueve la asociación entre una persona y un conjunto de datos.	DeId
5	Son datos generados artificialmente, como forma opuesta a recolectarlos del mundo real.	DS

Pregunta 4 - Bonus (6 pts)

Responda las siguientes consultas

a) (2 pts.) ¿Cuáles y en qué consisten las 3 propiedades fundamentales de los Sistemas Distribuidos?

Solución: Las 3 propiedades de los sistemas distribuidos son:

- C: Consistency (Consistencia), todos los usuarios ven lo mismo
- A: Availability (Disponibilidad), todas las consultas reciben una respuesta
- P: Partition Tolerance (Tolerancia al Particionamiento) el sistema sigue funcionando pese a estar físicamente dividido

b) (2 pts.) ¿Cuál de la siguiente opción es falsa con respecto a la BD Documentales?

I. Usan principalmente documentos JSON

II. Ideales para realizar cruce de datos complejos

III. Lenguaje de consulta poderoso

IV. No necesitan respetar un esquema de datos

Solución: La opción que es falsa es la II. Las BD documentales no están ideadas para realizar cruces de datos (joins) complejos. Esto debido a que esta lógica no la soportan y se debe realizar en el código de la aplicación.

c) (2 pts.) Si tengo una relación $R(A, B, C, D)$. Además $A \to B$, $A \to D$ y $D \to C$. ¿Cuál de las siguientes afirmaciones son verdaderas? Justifique su respuesta.

I. $A \to B, D$

II. $B \to D$

III. $A \to C$

IV. $B \to C$

Solución: Las afirmaciones que son verdaderas son:

- I es verdadero. Esto corresponde a la observación 4 de dependencias funcionales.
- II es falso. A partir de las dependencias del enunciado, no es posible establecer esta dependencia
- III es verdadero. Esto corresponde a la observación 1 de las dependencias funcionales
- IV es falso. A partir de las dependencias del enunciado, no es posible establecer esta dependencia.

Torpedo

Sintaxis consulta SQL

```
SELECT [ALL | DISTINCT] * | [COUNT, MAX, MIN, SUM, AVG](col1), col2, ...
       [AS alias_column]
FROM table1
     [JOIN table2 ON table1.common_field = table2.common_field]
WHERE condition1 [AND | OR condition2]
      [BETWEEN value1 AND value2]
      [IN (value1, value2, ...)]
      [LIKE pattern]
      [= | <> | <= | >= | < | >] value
GROUP BY column1, column2, ...
HAVING aggregate_condition [COUNT, MAX, MIN, SUM, AVG]
ORDER BY column1 [ASC | DESC]
UNION | INTERSECT | EXCEPT
[SELECT column1 FROM another_table]
LIMIT row_count;
```

Sintaxis Stored Procedure

```
CREATE OR REPLACE FUNCTION <nombre_función> (<argumentos>)
RETURNS [VOID | TABLE(...) ] AS $$
DECLARE
    <declaración_variables>
BEGIN
    <sentencias_SQL>
END; $$ language plpgsql
```

Sintaxis Trigger

```
CREATE TRIGGER <nombre_trigger>
[BEFORE | AFTER | INSTEAD OF] [ DELETE | UPDATE | INSERT] ON <tabla>
[FOR EACH ROW | FOR EACH STATEMENT]
BEGIN
    UPDATE <tabla_a_actualizar>
    SET <atributo> = <valor>
    WHERE <condición>
END;
```

Sintaxis PL/pgSQL

```
BEGIN
    IF <condición> THEN
        <cuerpo_if>
    ELSE
        <cuerpo_else>
    END IF;
    INSERT INTO <tabla> VALUES (<tupla_valores>);
    FOR <variable> IN <iterable>
```

```
    LOOP
        <cuerpo_loop>
    END LOOP;
    RETURN QUERY <consulta_SQL>;
END;
```