# Ranking Retrieval with PostgreSQL

Profesor Heider Sanchez

**Integrantes**

- Pedro Domínguez
- Eduardo Arróspide

El objetivo de este laboratorio es poner a prueba las técnicas de indexación de textos en PostgreSQL (full-text search index) mediante tres experimentos.

## P1. Sequential Scan vs GIN:

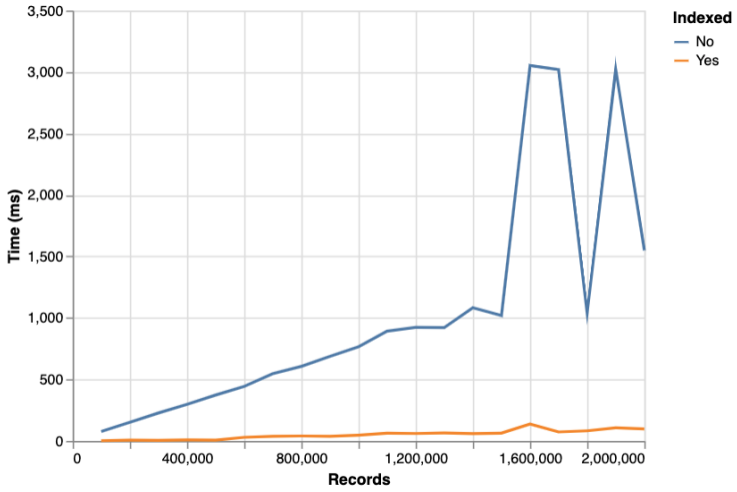
El primer experimento consiste en probar el índice invertido GIN representando el texto con **trigramas**. Un trigrama es un grupo de tres caracteres consecutivos tomados de una cadena. Ejemplo, los trigramas de la palabra “amor” son “amo” y “mor”. Indexar un atributo tipo texto con trigramas es eficaz en la mayoría de lenguajes naturales mejorando considerablemente las búsquedas textuales.

https://www.postgresql.org/docs/13/pgtrgm.html

Tomando como base el script dato en clase, se le pide realizar lo siguiente:

- Crear una tabla con dos atributos textuales, uno sin indexar y el otro indexado.
- Llenar datos aleatorios para diferentes cantidades.
- Ejecutar consultas sobre ambos atributos y tomar los tiempos

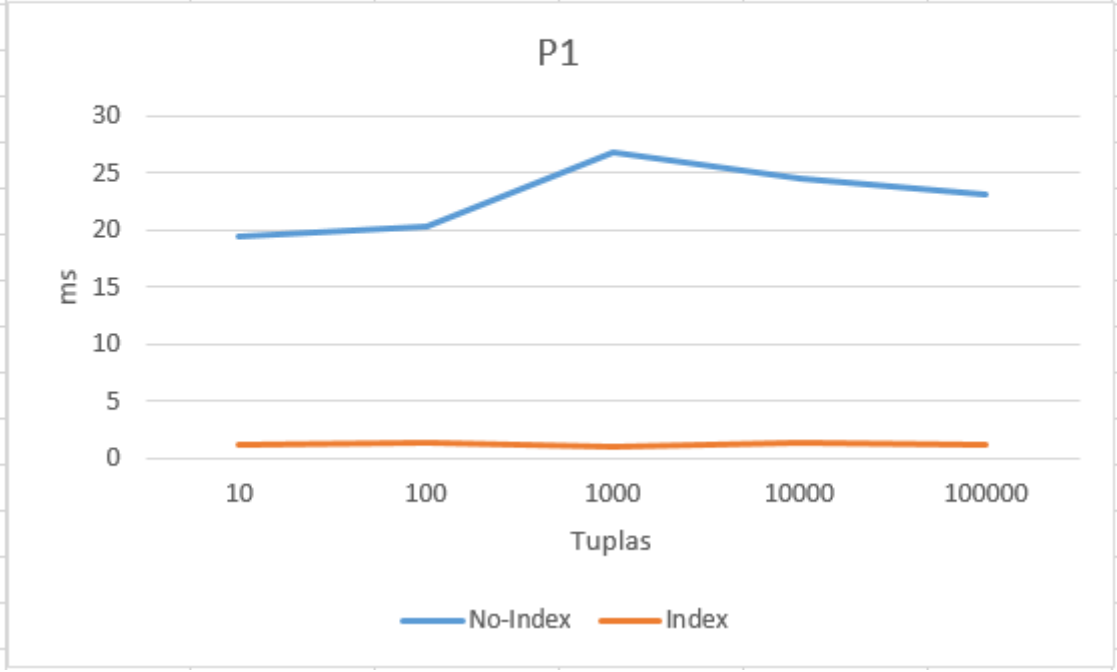
**Mostrar el plan de ejecución y un gráfico como resultado de la experimentación (ver gráfico de referencia).**



**Plan de ejecución**


**Gráfica de experimentación**

| | 10 | 100 | 1000 | 10000 | 100000 |
|---|---|---|---|---|---|
| **No-Index** | 19.38 | 20.324 | 26.788 | 24.564 | 23.156 |
| **Index** | 1.128 | 1.433 | 1.084 | 1.345 | 1.18 |



**P2. Full-text search on Films**

El segundo experimento consiste en aplicar el índice invertido GIN sobre los atributos textuales de la tabla “film” (dvdrental).

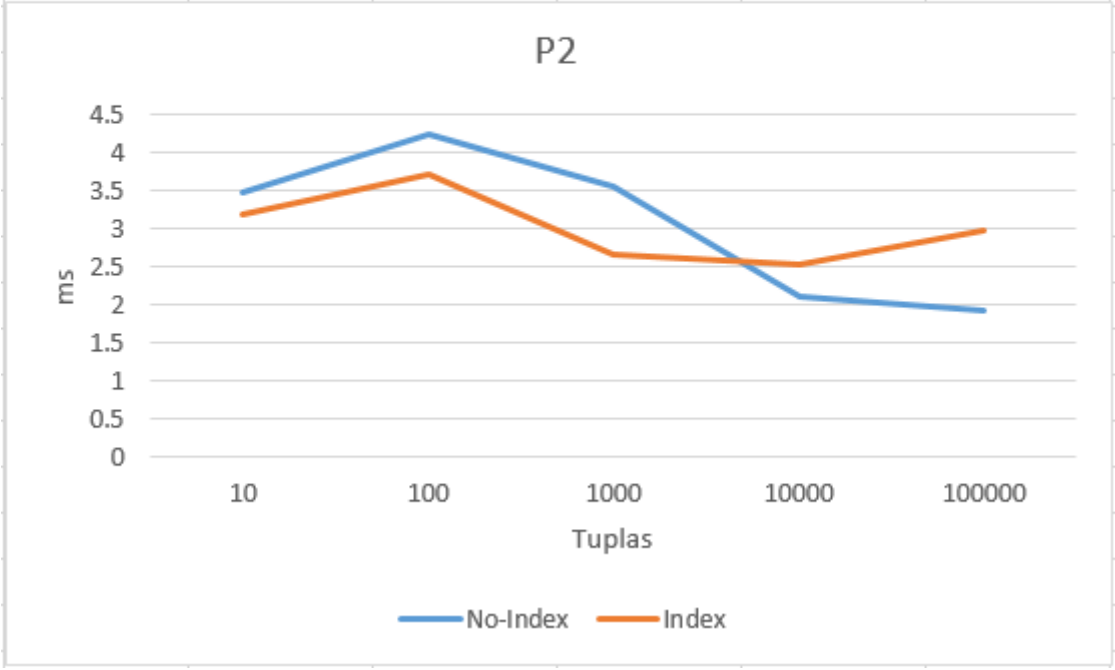
- Restaurar la base de datos en su servidor PostgreSQL
- Crear un nuevo atributo indexado compuesto por el titulo y la descripción de la película.
  - El tipo de dato corresponde al vector de pesos de los términos
- Ejecutar consultas sobre los atributos sin indexar y sobre el atributo indexado
  - Tomar los tiempos para diferentes rankings (top k)

**Mostrar el plan de ejecución y un gráfico como resultado de la experimentación**

**Plan de ejecución**


**Gráfica de experimentación**

| | 10 | 100 | 1000 | 10000 | 100000 |
|---|---|---|---|---|---|
| No-Index | 3.488 | 4.235 | 3.564 | 2.125 | 1.924 |
| Index | 3.189 | 3.72 | 2.657 | 2.529 | 2.992 |



**P3. Full-text search on News**

El tercer experimento consiste en aplicar el índice invertido GIN sobre los atributos textuales de la tabla “articles” (all the news).

- Crear la tabla Articles y llenar los datos desde los archivos CSV
- Crear un nuevo atributo indexado compuesto por el titulo y el contenido de la noticia.
  - El tipo de dato corresponde al vector de pesos de los términos
- Ejecutar consultas sobre los atributos sin indexar y sobre el atributo indexado
  - Tomar los tiempos para diferentes rankings (top k)

**Mostrar el plan de ejecución y un gráfico como resultado de la experimentación**

**Plan de ejecución**


**Gráfica de experimentación**

| | 10 | 100 | 1000 | 10000 | 100000 |
|---|---|---|---|---|---|
| No-Index | 0.579 | 0.872 | 0.905 | 0.77 | 0.928 |
| Index | 1.839 | 0.729 | 0.805 | 1.204 | 0.811 |