Islas Canarias
Del 15 al 19 de noviembre de 2021

Universidad de La Laguna

istac | INSTITUTO CANARIO DE ESTADÍSTICA

# XXI Jornadas de Estadística de las Comunidades Autónomas

## Diseño de una Base de Datos Orientada a Grafos para la gestión de datos en iDatos

Jesús Alberto González Yanes, ISTAC
Rafael Betancor Villalba, ISTAC
Luz Marina Moreno de Antonio, ULL
Isabel Sánchez Berriel, ULL

https://www.linkedin.com/company/istac/

@istac_es

# Índice:

- Introducción
- Objetivos
- Metodología
- Resultados
- Conclusión

# Introducción:

- Estadísticas multifuente:
  - Registros administrativos
    - padrón municipal, afiliados seguridad social, demandantes de empleo, DIRCE, …
  - Fuentes complementarias que ayudan a mejorar la calidad de los datos: APIS

# Introducción:

- *Una misma unidad de análisis, datos en distintas fuentes*
- Solución: **datos maestros** en iDatos:
  - Determinar un único elemento de referencia
  - Tablas que registran las relaciones entre los diferentes registros
- Implementación en una BDD relacional (PostgreSQL)

# Introducción:

- Dominios en iDatos:
  - población-hogares
  - empresas-establecimientos
  - lugares
  - edificios-viviendas-locales

# Introducción:

- Estructura:
  - **Directorios**: Registros conceptualmente relacionados
  - **Registros**: Ficheros relacionados con una unidad de análisis
  - **Capas**: ficheros de un registro según funcionalidad de integración

# Introducción:

- Variables:
  - **Nucleares:** Variables clave de un registro
  - **Normalizadas**: Variables que no siendo claves se quieren normalizar en iDatos
  - **Variables de entidades relacionadas:** Identificadores que vinculan unidades en dos registros

# Introducción:

- Ejemplo:
  - **IDT población:** Tabla de identificación única de una persona
  - **IDF directorio población:**
    - Variables Nucleares: *Variables identificadoras, Tipo de documento, Nº de identificación de la persona, letra de control, ..*
    - Variables Normalizadas: *Nombre completo, Primer nombre, Resto de nombres, Apellidos, Fecha de nacimiento, sexo, ...*

# Introducción:

- Tablas:
  - **DAT:** Ficheros de datos en un registro
  - **IDT:** Representantes únicos en un registro
    - Datos maestros en iDatos
  - **IDF:** Datos que complementan a un registro
    - Incluyen varias versiones
    - Variables nucleares y estandarizadas.

# Introducción:

- Tablas:
  - **URD:** Almacena relaciones con los datos maestros:
    - **IDT a IDF**
    - **$IDT_{D1}$ con $IDT_{D2}$**
    - **IDT – DAT**
  - Se recogen datos del tipo y calidad del enlace

# Introducción:

- Directorios:
  - ## Población y hogares:

    Padrón Municipal de Habitantes (PMH) (2003-2021); Demandantes de empleo según SISPE (2011-2021); Afiliaciones a la Seguridad Social (2011-2021); Afiliaciones a MUFACE (2011-2021); Contratos Laborales Registrados (2011-2021); Llamamientos Laborales (2011-2021); Egresados universitarios (2015-2018**)**

  - ## Calles y direcciones

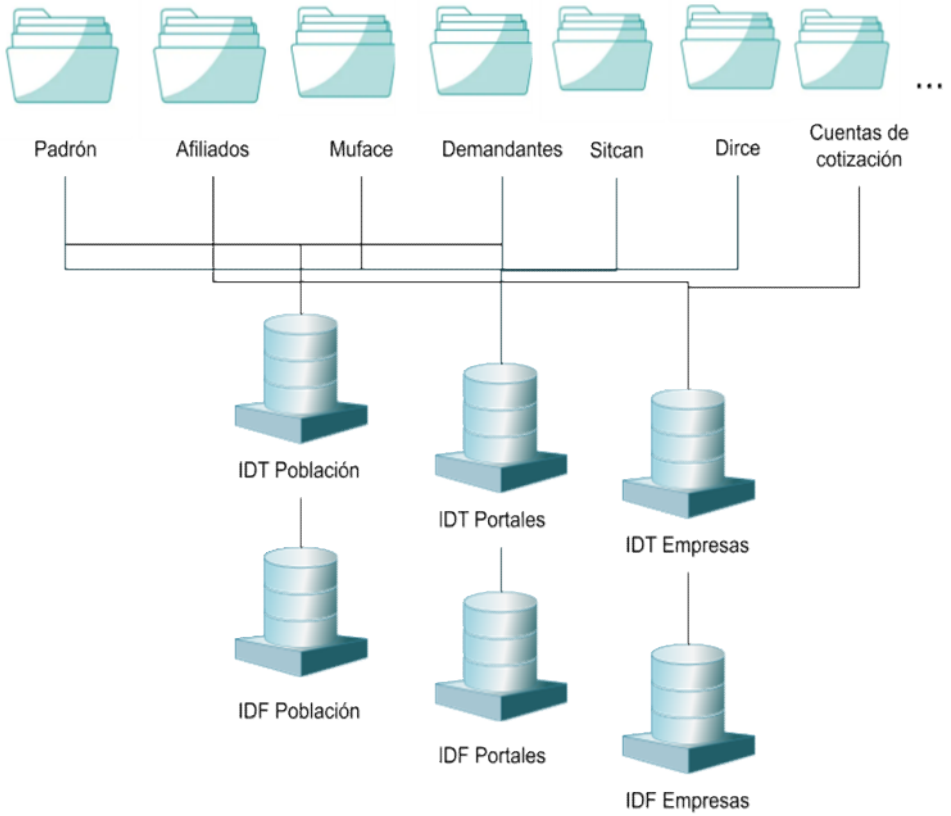
    Fichero de vías y portales del Sistema de Información Territorial de Canarias (SITCAN); Registros administrativos georreferenciadas con SITCAN y APIs (Google, Komoot, Cartociudad, etc); Padrón Municipal de Habitantes, demandantes de empleo, afiliaciones a la Seguridad Social, afiliaciones a MUFACE, contratos laborales, llamamientos laborales y cuentas de cotización

# Introducción:

- Directorios:
  - **Directorio de Unidades Económicas (DUE)**

    Registro Integrado de Empresas [de Canarias]; Trabajadores autónomos de afiliaciones en la Seguridad Social; Cuentas de Cotización a la Seguridad Social; DIRCE (Directorio Central de Empresas); Base de datos SABI; Inventario de Entes del Sector Público; Estadística Estructural de Empresas; Encuesta Industrial Anual de Productos; Empresas Inscritas en la Seguridad Social

# Arquitectura iDatos:

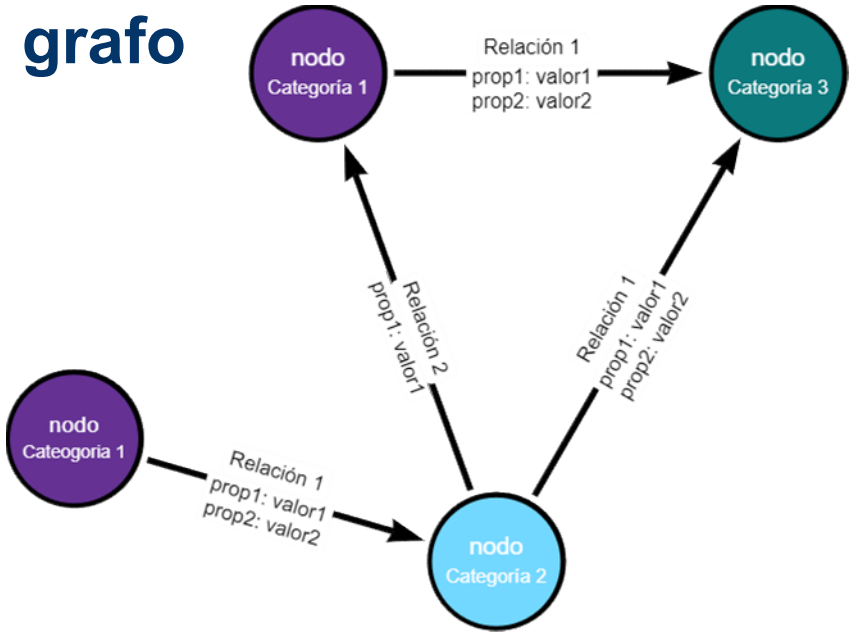


# Introducción:

- Caso de aplicación **EPA-Reg**:
  - Datos de la población activa a escala municipal y submunicipal
  - **Fuentes Principales**:
    - Padrón Municipal de Habitantes
    - Afiliaciones a la Seguridad Social o Muface
    - Demandas en el Servicio Canario de Empleo.

# Introducción:

- BDD orientadas a Grafos (BDOG):
  - Modelo de datos un **grafo**



# Introducción:

- BDD orientadas a Grafos:
  - Ventajas en **conjuntos de datos conectados** frente al uso intensivo de JOINS:
    - Procesamiento nativo de grafos: **índices libres de adyacencia**
    - **Rendimiento en consultas con varios saltos permanece constante**

# Introducción:

- BDOG para datos maestros:
  - Conexión entre diferentes silos de datos.
  - Representación de relaciones, jerarquías y metadatos.
  - Complementan sistemas existentes.
  - Simplicidad del lenguaje de consultas
  - Facilidad de visualizar y comunicar
  - Algoritmos de aprendizaje

# Introducción:

- Neo4j:
  - Modelo de datos un grafo
  - Procesamiento nativo de grafos, clusters distribuídos
  - Altamente escalable
  - Transacciones ACID implementada en Java y Scala

# Introducción:

- Neo4j:
  - Herramientas visuales que facilitan la consulta y administración de los datos.
  - APIs RESTful y Java entre otros.
  - Lenguaje de consulta Cypher
  - Amplia comunidad y documentación disponible.

# Objetivos:

- **Diseño del grafo** para los datos maestros de iDatos
- **Implementación** del grafo en **Neo4j**
- **Muestra** que simula **iDatos anonimizado**
- Análisis de **rendimiento de consultas**:
  - Relacionadas con **EPA-Reg**
  - **Relaciones inferidas** sobre el grafo.

# Metodología :

- **Muestra de datos anonimizados** en iDatos, junio y septiembre del **año 2017**:
  - Semilla aleatoria en las **tablas IDT** de cada registros.
  - **Casos vinculados** en el conjunto de **tablas IDF y DAT** obtenidos de las tablas URD
  - Ampliación de la muestra forzando la inclusión de casos de **relaciones transitivas**.

# Metodología :

- Relaciones directas:
  - $IDT_{DIRECTORIO\_A}$-$IDF_{DIRECTORIO\_A}$
  - $IDT_{DIRECTORIO\_A}$-Datos $_{DIRECTORIO\ A}$
- Relaciones transitivas:
  - Datos Raw $_{DIRECTORIO\ A}$-$\mathbf{IDT_{DIRECTORIO\_A}}$
  - $\mathbf{IDT_{DIRECTORIO\_A}}$- $\mathbf{Datos_{DIRECTORIO\_B}}$
  - $\mathbf{Datos_{DIRECTORIO\_B}}$ - $IDT_{DIRECTORIO\_B}$

# Introducción:

- **Datos en la muestra:**

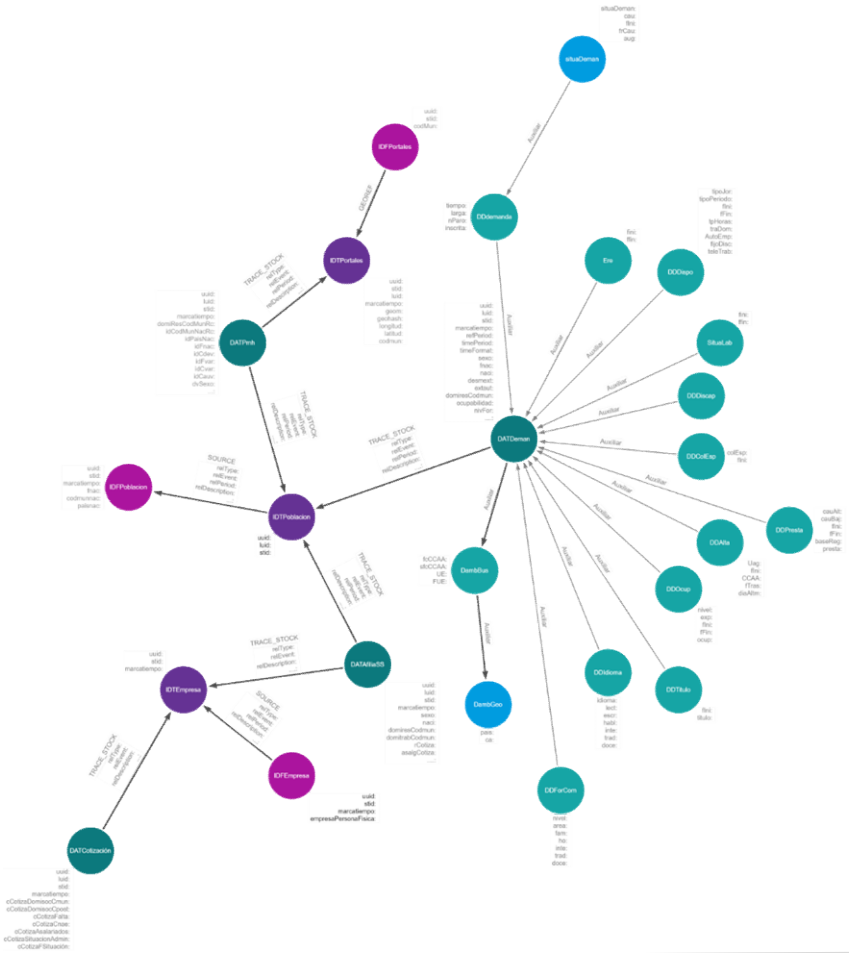
| Tabla iDatos | |
|---|---|
| IDT Empresas | 154 |
| IDF Empresas | 982 |
| URD | 16.650 |
| IDT Población y Hogares | 2.472 |
| IDF Población y Hogares | 4.826 |
| URD Población | 169.926 |
| IDT Portales | 2.126 |
| IDF Portales | 3.896 |
| URD Portales | 102.624 |

| Registro Administrativo | |
|---|---|
| CCotiza Junio | 89.558 |
| CCotiza Septiembre | 89.193 |
| Afiliados Seguridad Social Junio | 2.216 |
| Afiliados Seguridad Social Septiembre | 2.586 |
| Demandantes de empleo Junio | 273 |
| Demandantes de empleo Septiembre | 153 |
| Padrón 2017 | 2.465 |
| | |
| | |

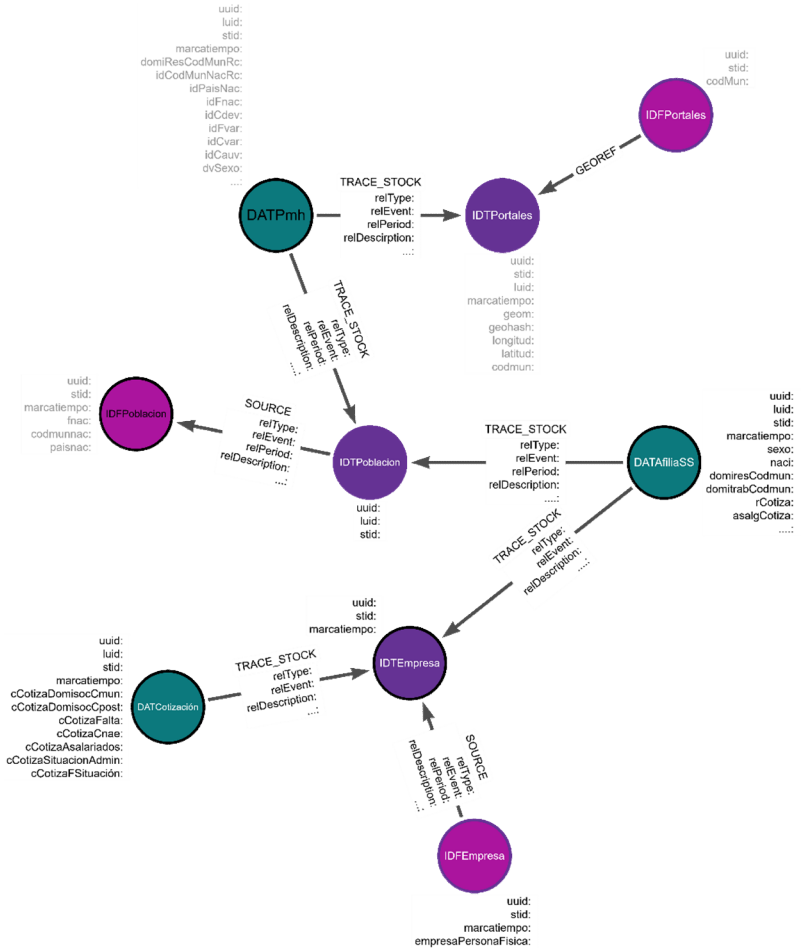
# Diseño del Grafo :

- **Nodos**: Datos en IDT, IDF y Datos crudos.
- **Etiquetas**: Indican el directorio
- **Propiedades**: Atributos en las tablas relacionales
- **Aristas**: Relaciones URD
- **Etiqueta de la relación**: tipo de relación URD
- **Nodos auxiliares**: Atributos Multivaluados.

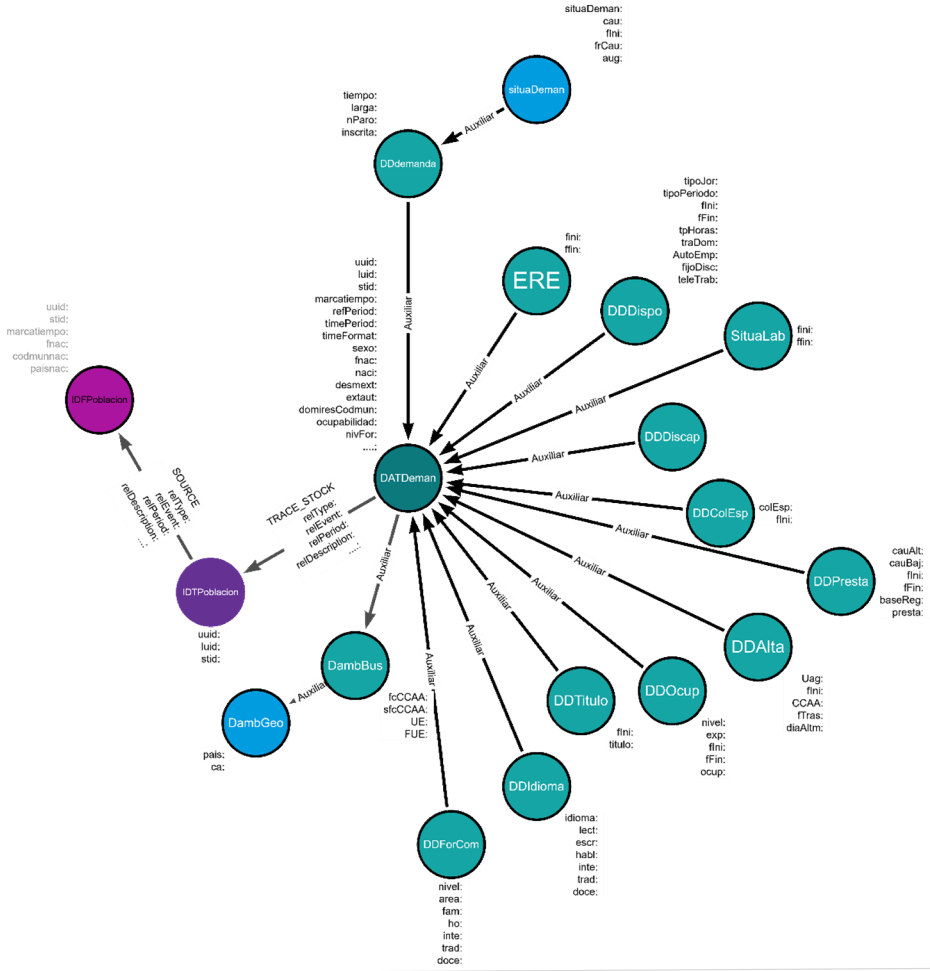
# Diseño del Grafo:



# Diseño del Grafo:



# Diseño del Grafo:



# BDD Relacional :

- Tablas IDT, IDF de los tres directorios:

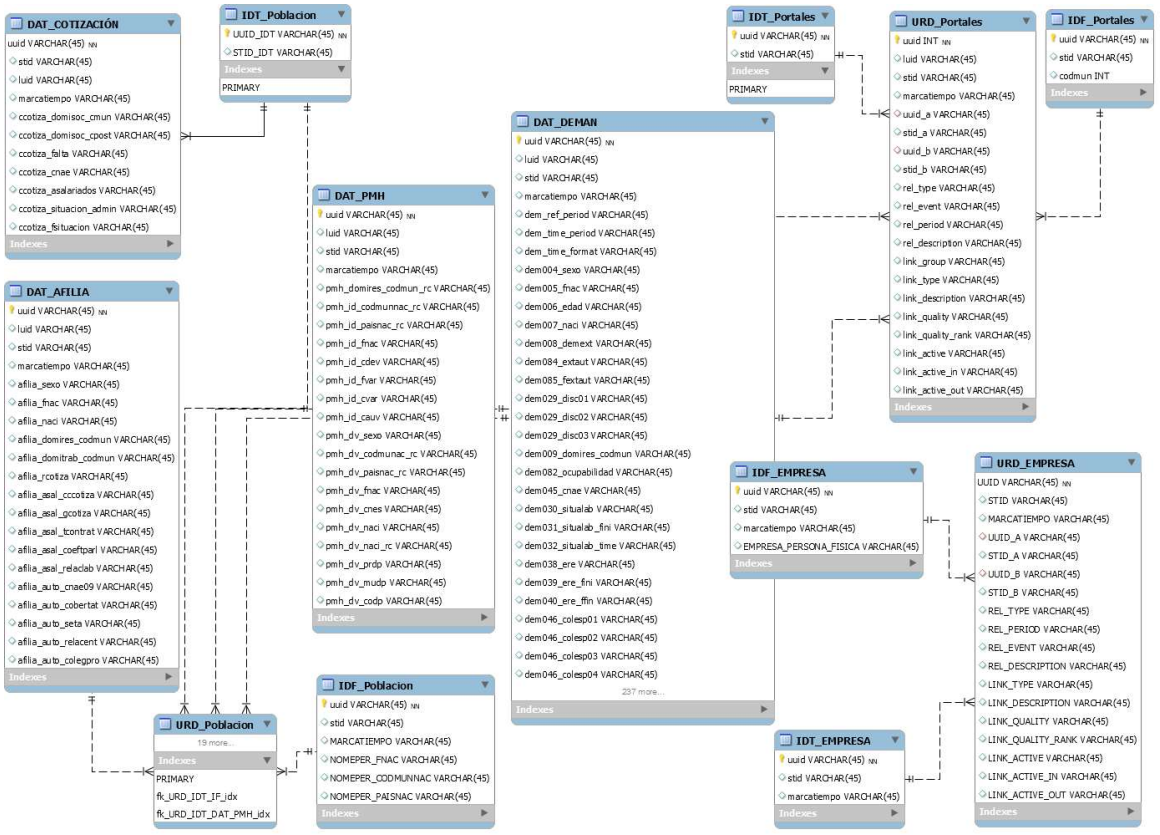
Población y Hogares, Calles y Direcciones, Unidades Económicas.

- Tablas por cada fuente:

Afiliados a la Seguridad Social, Demandantes de Empleo (junio y septiembre de 2017), Padrón Municipal de Habitantes (julio de 2017 ), Cuentas de Cotización a la Seguridad Social (junio y septiembre de 2017).

- Tablas URD de cada directorio:
  - Relaciones entre las fuentes y los datos maestros
  - Relaciones IDT e IDF

# Base de datos relacional:



# Resultados:

- Estudio comparativo del rendimiento en ambos sistemas:
  - Consultas complejas
  - Consultas habituales en EPA-Reg
  - Dos ejecuciones con cada uno de los sistemas gestores, mismo día y misma sesión de usuario
  - Máquina virtual del servicio iaaS de la ULL

    S.O. Ubuntu, 16 Gigas de RAM y 20 Gigas de disco; PostgreSQL; Neo4j

# Resultados:

- Empresas que han aumentado el número de empleados y en qué cantidad de junio a septiembre 2017
  - **URD_EMPRESAS; IDT_EMPRESAS; CUENTAS_COTIZACIÓN**

```
match(n:idt_empresa)-[:TRACESTOCK]-(:dat_ccotiza)
call {
     with (n)
     match (n)-[r1:TRACESTOCK]-(q:dat_ccotiza{stid:"rd_cotizacion_segsoc.dat_ccotiza20170930_v01"})
     return sum(tointeger(q.ccotiza_asalariados)) AS num1
}
call {
     with (n)
     match (n)-[r1:TRACESTOCK]-(q:dat_ccotiza{stid:"rd_cotizacion_segsoc.dat_ccotiza20170630_v01"})
     return sum(tointeger(q.ccotiza_asalariados)) AS num2
}
with distinct n, num1, num2
where num1>num2
return distinct n.uuid, num1-num2 AS diferencia;
```

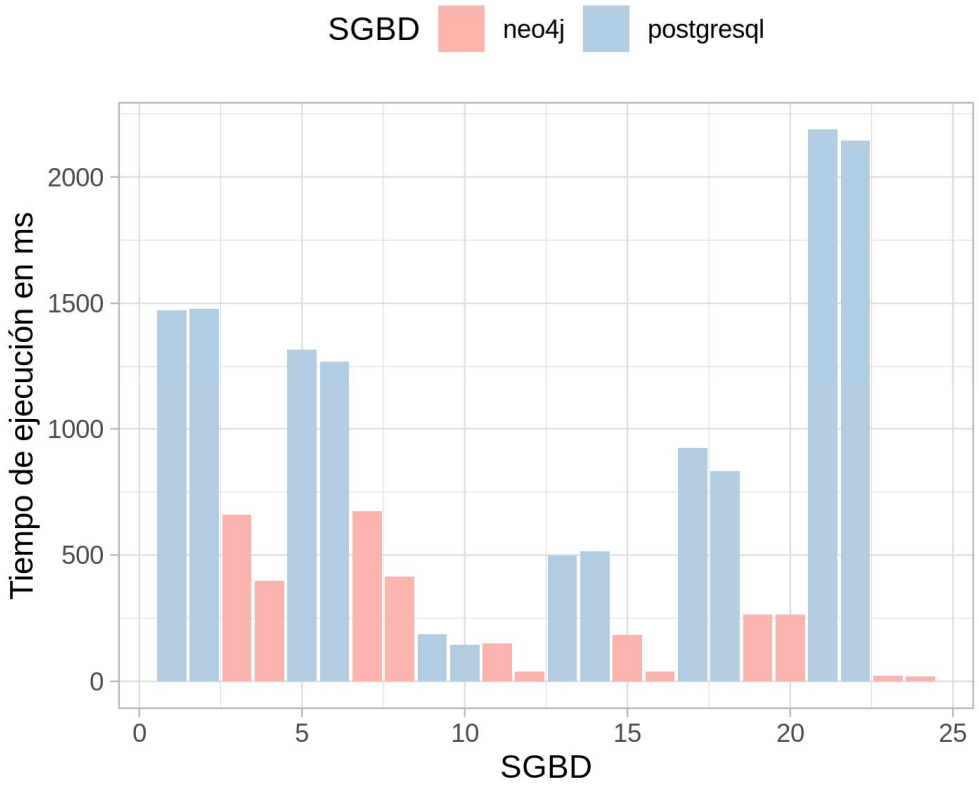
# Resultados:

- Consultas habituales en EPA-Reg:
  - Personas en archivos de **afiliados** con **valores diferentes de sexo y fecha de nacimiento**
    - **IDT_POBLACIÓN; URD_POBLACIÓN; DAT_AFILIADOS**
  - Personas con valor **sexo diferente** en alguna de las tablas **IDF, Afiliados, Demandante y Padrón**
    - **IDT_POBLACIÓN; IDF_POBLACIÓN; DAT_AFILIADOS; DAT_DEMANDANTES; DAT_PMH**

# Resultados:

- Consultas habituales en EPA-Reg:
  - Personas con valor **sexo diferente** en alguna de las tablas **IDF, Afiliados, Demandante y Padrón**
    - IDT_POBLACIÓN; IDF_POBLACIÓN; DAT_AFILIADOS; DAT_DEMANDANTES; DAT_PMH
  - Listado de personas que están en alguno de los ficheros de **afiliados o demandantes y no están en PMH**
    - IDT_POBLACIÓN; IDF_POBLACIÓN; URD_POBLACIÓN; DAT_AFILIADOS; DAT_DEMANDANTES; DAT_PMH

# Resultados:


SGBD
neo4j
postgresql
Tiempo de ejecución en ms
2000
1500
1000
500
0
0
5
10
15
20
25
SGBD

# Resultados:

| Consulta | Nodos | Exceso |
|---|---|---|
| Empresas que han aumentado el número de empleados y en qué cantidad | 37 | 2,793755913 |
| Empresas que han disminuido el número de empleados y en qué cantidad | 25 | 2,378676471 |
| Listado de personas con valores diferentes de sexo y fecha de nacimiento. | 173 | 1,755319149 |
| Personas en archivos de afiliados con valores diferentes de sexo y fecha de nacimiento | 3 | 4,572072072 |
| Personas con valor sexo en alguna de las tablas IDF, Afiliados, Demandante y Padrón diferente | 10 | 3,336501901 |
| Listado de personas que están en alguno de los ficheros de Afiliados, Demandantes, Cuentas de Cotización y no están en PMH | 9 | **100,6976744** |

# Conclusiones:

- Neo4j una alternativa adecuada a PostgreSQL para la gestión de datos maestros en iDatos
- Diseño fácilmente **adaptable** a medida que la infraestructura iDatos escale.
- **Disminuye el tiempo de ejecución,** especialmente en consultas que implican mayor cantidad de tablas.
- **Alta expresividad** que facilita la compresión del modelo y la sintaxis de las consultas.

Islas Canarias
Del 15 al 19 de noviembre de 2021

INSTITUTO CANARIO DE ESTADÍSTICA

# XXI Jornadas de Estadística de las Comunidades Autónomas

## Diseño de una Base de Datos Orientada a Grafos para la gestión de datos en iDatos

Jesús Alberto González Yanes, ISTAC
Rafael Betancor Villalba, ISTAC
Luz Marina Moreno de Antonio, ULL
Isabel Sánchez Berriel, ULL

**¡¡¡ GRACIAS !!!**

https://www.linkedin.com/company/istac/

@istac_es

Islas Canarias
Del 15 al 19 de noviembre de 2021

# ¡Gracias!

istac | INSTITUTO CANARIO DE ESTADÍSTICA

Universidad de La Laguna

https://www.linkedin.com/company/istac/

@istac_es