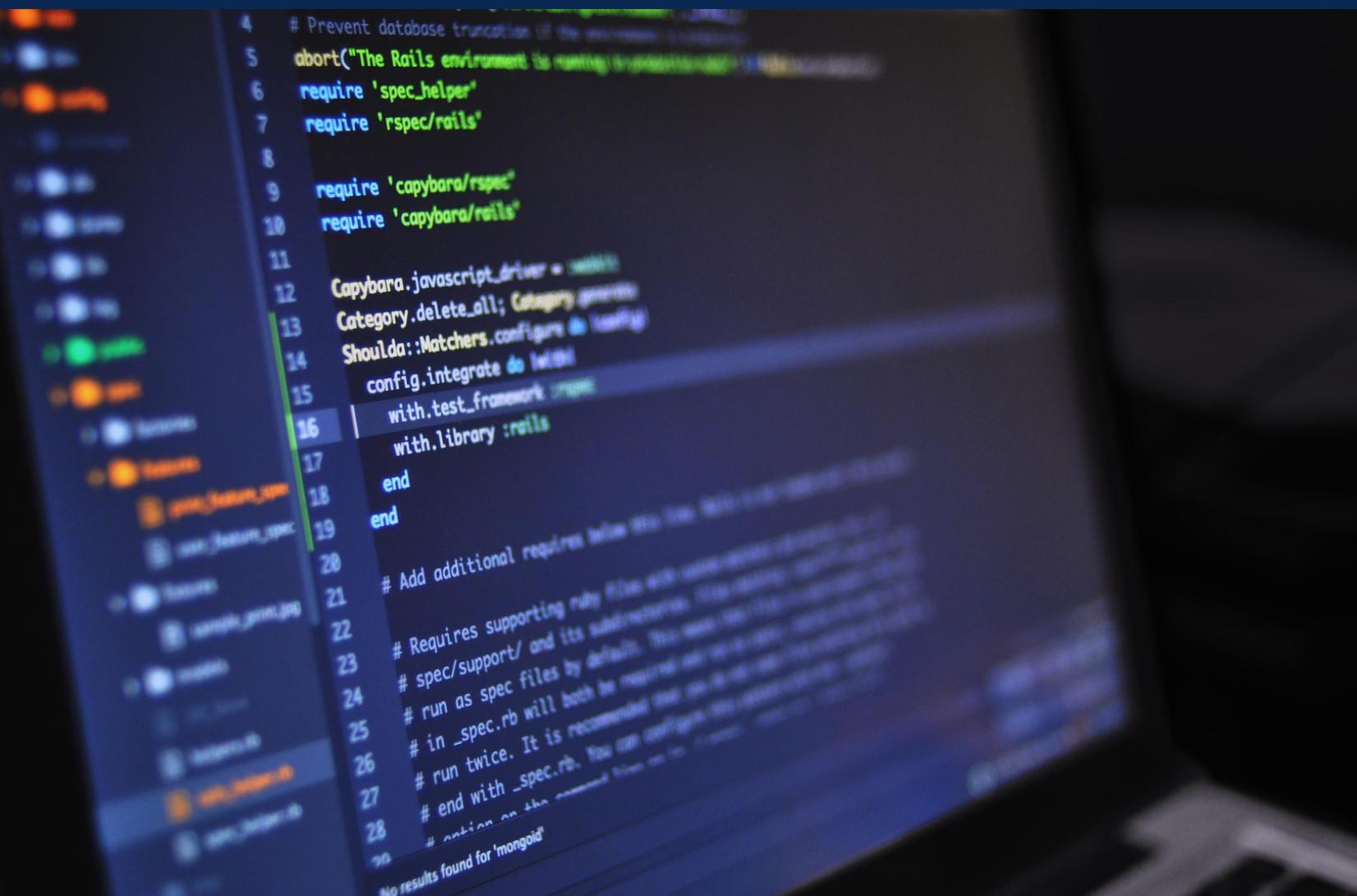


PRESENTACION FINAL PRÁCTICA 2024

JOSÉ GUILLERMO SEPÚLVEDA SALAZAR
VIERNES 15 DE MARZO DEL 2024

INTRODUCCIÓN



SE ESTABLECE UN MARCO PARA FORTALECER LOS PROCESOS ESTADÍSTICOS DE LA OFICINA DE ESTUDIOS Y ESTADÍSTICA (OEE), PARTICULARMENTE EN LA UNIDAD DE ESTADÍSTICAS, EL ENFOQUE SE CENTRA EN LA ACTUALIZACIÓN Y OPTIMIZACIÓN DE LOS PROCESOS DE RECOPILACIÓN Y ANÁLISIS DE DATOS, ASEGURANDO LA ENTREGA DE INFORMACIÓN PRECISA Y RELEVANTE. ESTE PLAN CONTEMPLA EL USO DE TECNOLOGÍAS ACTUALIZADAS Y METODOLOGÍAS EFICIENTES Y SEGURAS PARA LA GESTIÓN DE DATOS, PARA LA TOMA DE DECISIONES INFORMADAS DENTRO DEL SECTOR.

OBJETIVO GENERAL



Contribuir al fortalecimiento de los productos elaborados por la Unidad de Estadística, con el fin de proporcionar a la Oficina y al Ministerio datos confiables y actualizados del entorno CTCI.

OBJETIVOS ESPECÍFICOS

```
from selenium import webdriver
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from webdriver_manager.chrome import ChromeDriverManager
import time

# Configuración de Selenium con ChromeDriver
service = Service(ChromeDriverManager().install())
driver = webdriver.Chrome(service=service)

# URL que queremos hacer scraping
url = "https://data-explorer.oecd.org/vis?pg=0&bp=true&snb=6&tm=science&vw=tb&df[ds]=dsDisseminateFinalDMZ&df[id]=DSD_MST"

# Abrir la URL con Selenium
driver.get(url)

# Extraer el título, esperando a que los datos se estabilicen
try:
    WebDriverWait(driver, 10).until(EC.presence_of_element_located((By.CSS_SELECTOR, "h1.MuiTypography-root")))
    ante_titulo = ""
    while True:
        titulo = driver.find_element(By.CSS_SELECTOR, "h1.MuiTypography-root").text
        if titulo == ante_titulo:
            break
        ante_titulo = titulo
        time.sleep(1) # Espera un segundo antes de verificar nuevamente

    print("El título es:", titulo)
except Exception as e:
    print("No se pudo encontrar el título:", e)

# Extraer tabla
try:
    WebDriverWait(driver, 10).until(EC.presence_of_element_located((By.CSS_SELECTOR, "table.MuiTable-root")))

    # Encuentra la tabla
    tabla = driver.find_element(By.CSS_SELECTOR, "table.MuiTable-root")
    print('vamos bien por aki')

    # Extraer todas las filas de la tabla
    filas = tabla.find_elements(By.TAG_NAME, "tr")

    # Iterar sobre cada fila para extraer los datos de las celdas
    for fila in filas:
        celdas = fila.find_elements(By.TAG_NAME, "td")
        datos_fila = [celda.text for celda in celdas]
        print(datos_fila)

except Exception as e:
    print("Error al extraer datos de la tabla:", e)

# Cerrar el navegador
driver.quit()
```

- * **OE-1:** Actualizar indicadores de encuestas y datos administrativos asociados a CTCI en la plataforma Observa.
- * **OE-2:** Diseñar y proponer un diagrama de Entidad-Relación detallado para I+D a la Oficina de Estudio y Estadística.
- * **OE-3:** Implementar un sistema automatizado de extracción de información en la web a indicadores establecidos.

N E C E S I D A D E S I D E N T I F I C A D A S



LA TAREA PRINCIPAL DEL PRACTICANTE CONSISTE EN ACTUALIZAR LOS DATOS DE LAS ENCUESTAS DE I+D DEL 2021 Y SIES DE LOS AÑOS 2022 Y 2023, ABORDANDO LA NECESIDAD DE MEJORAR LA COHESIÓN Y REDUCIR LA REDUNDANCIA EN LAS BASES DE DATOS SQL ACTUALES PARA ESTABLECER UN SISTEMA MÁS EFICIENTE EN EL FUTURO. ADEMÁS, SE IDENTIFICARON PROCESOS MANUALES POCO EFICIENTES A FALTA DE DOCUMENTACIÓN, PROponiendo LA AUTOMATIZACIÓN PARA AGILIZAR LA RECOPILACIÓN DE DATOS. EL DESARROLLO SE LLEVARÁ A CABO EN PYTHON 3.9, UTILIZANDO MARKDOWN PARA DOCUMENTACIÓN Y MINICONDA PARA LA GESTIÓN DE LIBRERÍAS, TODO DENTRO DE JUPYTER NOTEBOOK.

RECOMENDACIONES

ANÁLISIS DE
INDICADORES I+D Y SIES

MANTENIMIENTO DE
DATOS OCDE

OBSERVACIONES TABLAS
DE INDICADORES I+D

A CONTINUACIÓN, SE PRESENTAN LOS PRODUCTOS QUE DERIVAN DEL TRABAJO DE PRÁCTICA DURANTE ESTE PERÍODO. ESTAS RECOMENDACIONES SON SEPARADAS POR LAS PRINCIPALES ACTIVIDADES REALIZADAS.

Jose-Sepulveda-Salazar/ Minciencia

Repository para realizar avances de mi práctica



1 Contributor 0 Issues 1 Star 0 Forks

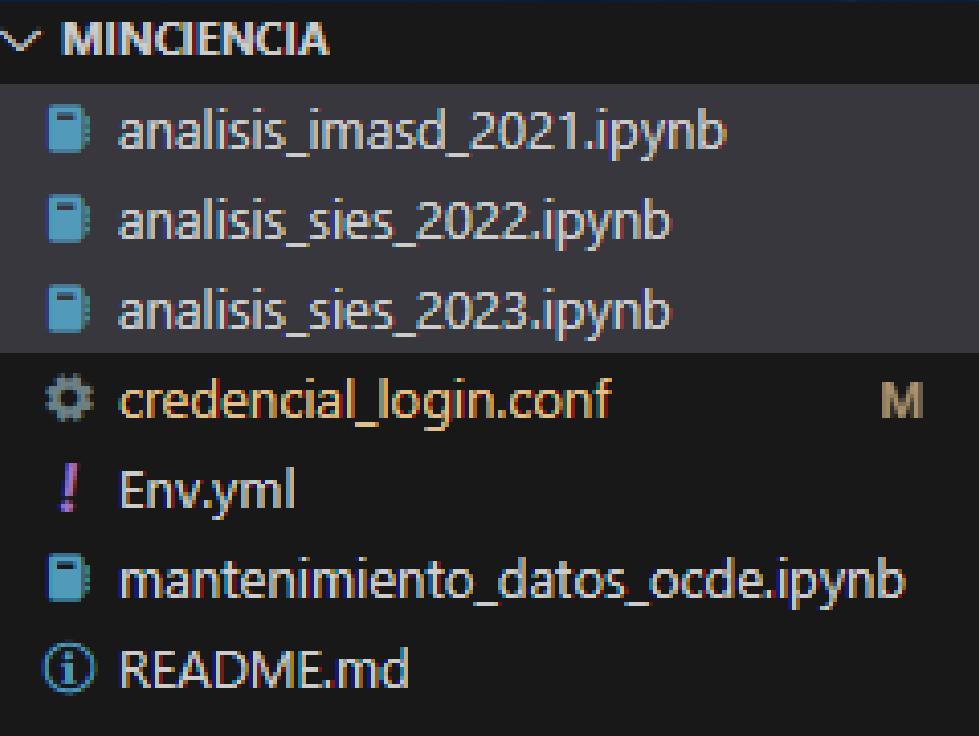
Jose-Sepulveda-Salazar/Minciencia: Repository para realizar avances de mi práctica

Repository para realizar avances de mi práctica. Contribute to Jose-Sepulveda-Salazar/Minciencia development by creating an account on GitHub.

[GitHub](#)

ANÁLISIS DE INDICADORES I+D Y SIES

La automatización de indicadores mediante Jupyter abarca desde la recolección hasta el almacenamiento de datos, con validaciones rigurosas para su correcta integración. Se enfatiza la seguridad mediante mecanismos de autenticación, alineados con estándares de protección de datos. El proceso se divide en extracción, transformación y carga (ETC), incluyendo verificaciones de actualización y técnicas como Web Scraping y conexiones API. Para prevenir errores, se implementan capas adicionales antes de la carga final de datos, requiriendo validación manual para asegurar precisión e integridad, reforzando así la gobernanza de datos y la fiabilidad del sistema.



MANTENIMIENTO DE DATOS OCDE

El sistema automatizado extrae datos de la OCDE utilizando una interfaz API y los analiza comparando tres tablas específicas en el SGBD: "investigadores_1000_long", "gastoid_porcentaje_pib_long" y "Investigadores_Mujeres_OCDE", usando DataFrames para el análisis. Identifica discrepancias significativas, clasificándolas en tres categorías para una interpretación estructurada. Los resultados se exportan a un archivo Excel con un sistema de versionado por fecha y hora, mejorando la organización y trazabilidad, complementado por un resumen cuantitativo para una revisión rápida y efectiva del análisis realizado.

MINCIENCIA

- analysis_imasd_2021.ipynb
- analysis_sies_2022.ipynb
- analysis_sies_2023.ipynb
- credencial_login.conf
- Env.yml
- mantenimiento_datos_ocde.ipynb
- README.md

OBSERVACIONES TABLAS

DE INDICADORES I + D

OCDE-promedio-OCDE-UE	Investigadores_Mujeres_OCDE	Gastoid_Matriz
+ country: varchar(255) + Year: int(4) + stem: float(8) + nombre_esp: varchar(255) + prom_Invest1000: float(8) + prom_gastoid: float(8)	+ País Inglés: text + País Español: text + Año: text + Porcentaje de mujeres investigadoras respecto al total de inves: float(8)	+ agno: float(8) + financiamiento: varchar(255) + ejecucion: varchar(255) + monto: float(8)
Ley_ID_Intramuro	PersonalID_Nivel_Titulacion_Sector	gastoid_acteco
+ anho: float(8) + ejecuta_leyid: text + id_intramuro: float(8)	+ anho: int(4) + UNIDAD_DECLARANTE: text + doctorado: float(8) + magister: float(8) + profesional: float(8) + tecnicos: float(8) + otros: float(8) + Total: float(8)	+ anho: int(4) + a: float(8) + b: float(8) + c: float(8) + c20: float(8) + c21: float(8) + d: float(8) + e: float(8) + f: float(8) + g: float(8) + h: float(8) + i: float(8) + j: float(8) + k: float(8) + l: float(8) + m: float(8) + m70: float(8) + m71: float(8) + m72: float(8) + n: float(8) + o: float(8) + p: float(8) + q: float(8) + r: float(8) + s: float(8)
Personal_JCE_Sector	investigadores_1000_long	Personal_JCE_Sector_Investigadores
+ Año: float(8) + Sector de ejecución: text + sexo: text + Personal en I+D (JCE): float(8)	+ año: int(4) + paises: varchar(255) + valor: varchar	+ Año: float(8) + Sector de ejecución: text + Sexo: text + Investigadores (JCE): float(8)
gastoid_porcentaje_pib_long	empresas_n_empleados_n_investig_jce	investigadores_jce_sexo_macrozona
+ año: int(4) + país: varchar(255) + gastoid: float(8)	+ año: int(4) + sexo: varchar(255) + Tipo empresa: varchar(255) + Cantidad de empleados (JCE): float(8) + Investigadores (JCE): float(8)	+ año: int(4) + macrozona: varchar(255) + sexo: varchar(255) + Personal en ID JCE: float(8)
		Gasto_id_050606_resumida
		+ id_registro: int(4) + anho: int(4) + macrozona: varchar(255) + gasto ID como porcentaje PIB: float(8) + JCE por cada 1000 personas: float(8)

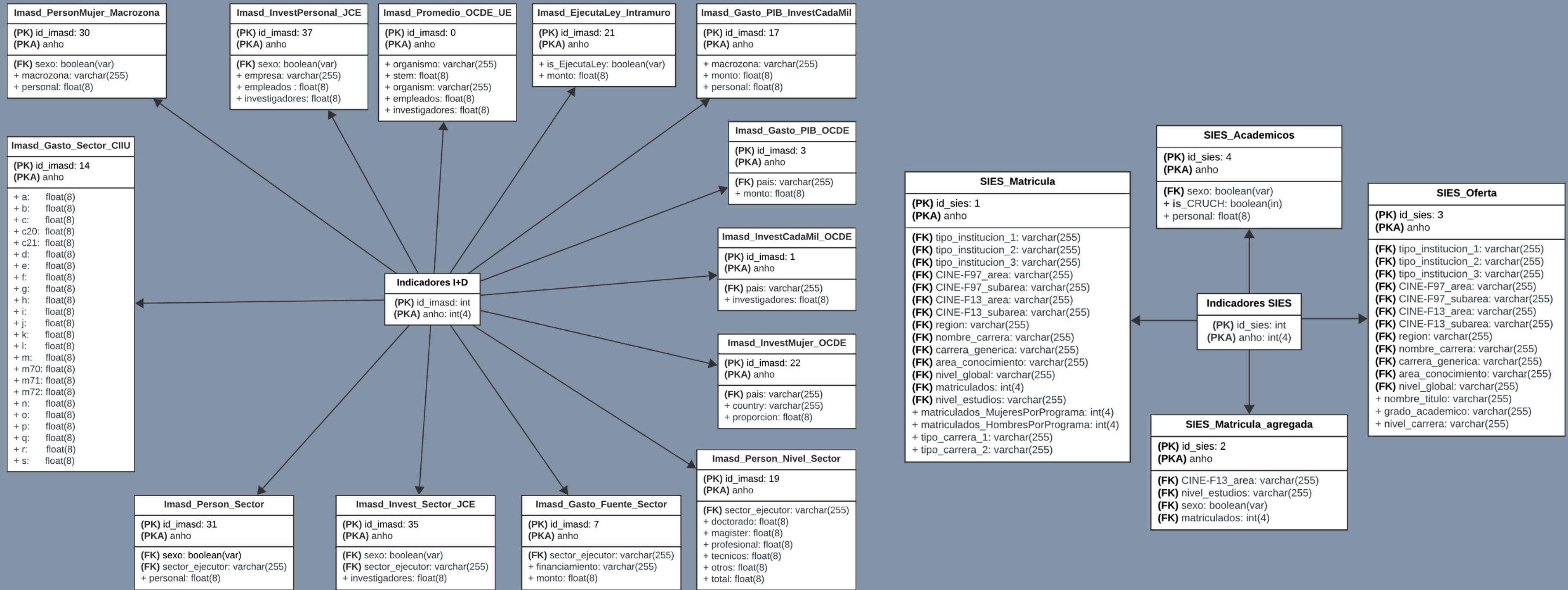
PROUESTA MODELO
ENTIDAD-RELACIÓN

ETIQUETADO DE TABLAS

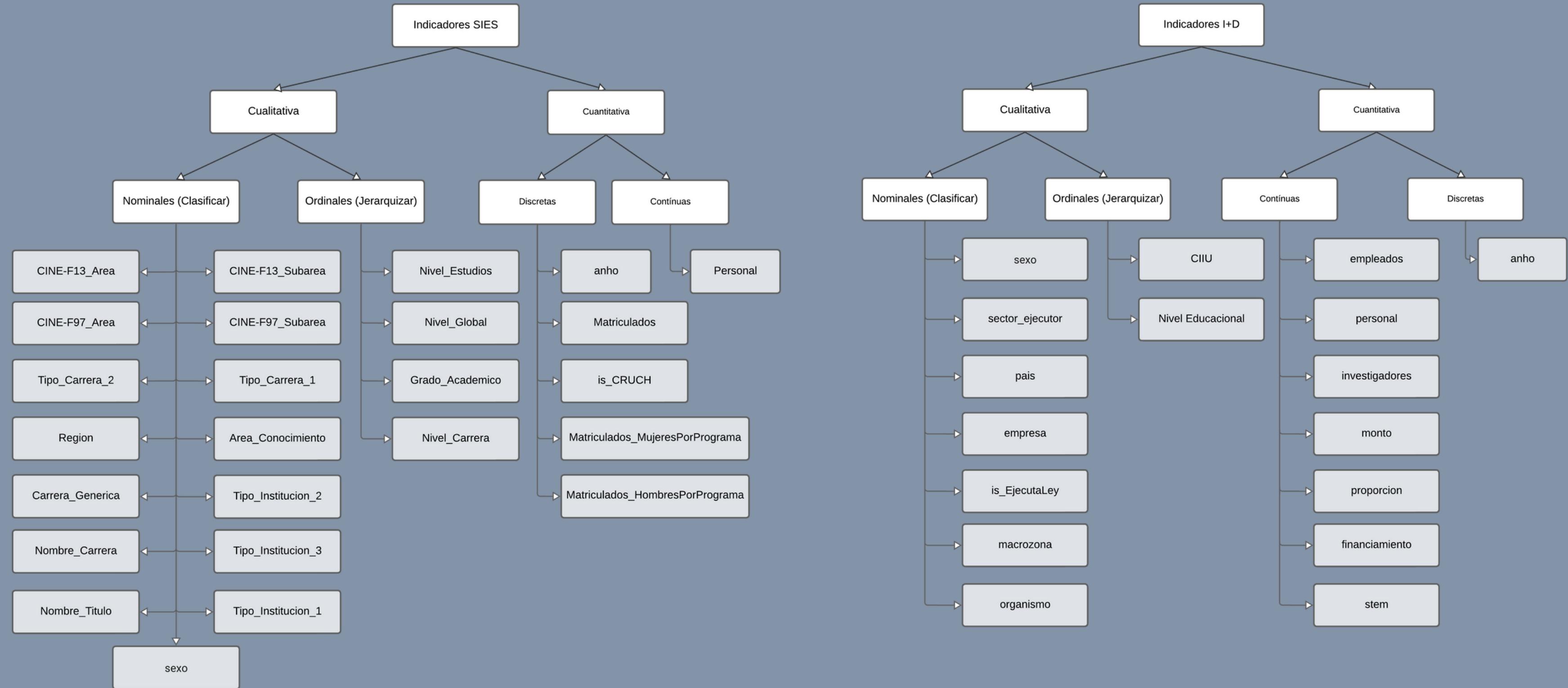
Original	Propuesta	Indicador
OCDE-promedio-OCDE-UE	imasd_Promedio_OCDE_UE	Promedio OCDE y Unión Europea.
investigadores_1000_long	imasd_InvestCadaMil_OCDE	Cantidad de investigadores(as) cada mil personas trabajando en países de la OCDE.
gastoid_porcentaje_pib_long	imasd_Gasto_PIB_OCDE	Gasto en I+D respecto al PIB en países de la OCDE.
Gastold_Matriz	imasd_Gasto_Fuente_Sector	Distribución del gasto en I+D según fuente de financiamiento y sector de ejecución.
gastoid_acteco	imasd_Gasto_Sector_CIIU	Gasto en I+D por empresas según su sector económico (clasificación CIIU rev.4)
Gasto_id_050606_resumida	imasd_Gasto_PIB_InvestCadaMil	Gasto en I+D como porcentaje del PIB e investigadores/as cada 1000 personas trabajando por macrozona.
PersonalID_Nivel_Titulacion_Sector	imasd_Person_Nivel_Sector	Distribución del personal en I+D según nivel educacional y sector de ejecución.
Ley_ID_Intramuro	imasd_EjecutaLey_Intramuro	Porcentaje de empresas que ejecuta montos bajo la Ley I+D entre las que realizan I+D intramuro.
Investigadores_Mujeres_OCDE	imasd_InvestMujer_OCDE	Porcentaje de investigadoras mujeres en países de la OCDE.
investigadores_jce_sexo_macrozona	imasd_PersonMujer_Macrozona	Porcentaje del personal dedicado a I+D que es mujer respecto al total de cada macrozona.
Personal_JCE_Sector	imasd_Person_Sector	Distribución del personal dedicado en I+D según su sexo por cada sector de ejecución.
Personal_JCE_Sector_Investigadores	imasd_Invest_Sector_JCE	Distribución por sexo de las jornadas completas equivalentes (JCE) trabajadas en I+D por investigadores según sector de ejecución.
empresas_n_empleados_n_investig_jce	imasd_InvestPersonal_JCE	Porcentaje de investigadores(as) que son mujeres en empresas beneficiadas por la Ley I+D. y Porcentaje de personal en I+D que son mujeres en empresas beneficiadas por la Ley I+D.

INDICADOR_CRITERIO_CARACTERÍSTICA RELEVANTE

PROUESTA BD

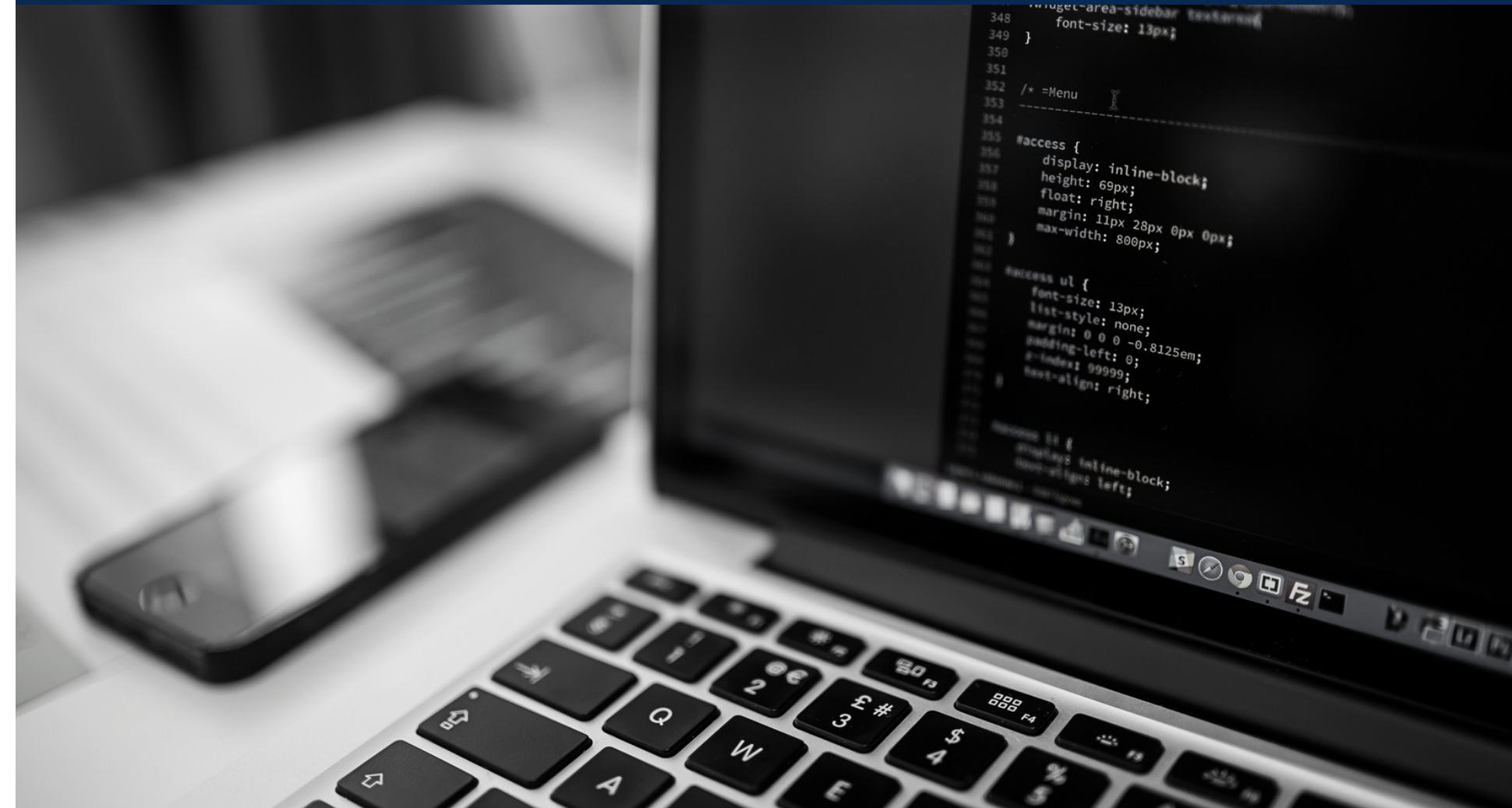


JUSTIFICACIÓN POR TIPO DE VARIABLE



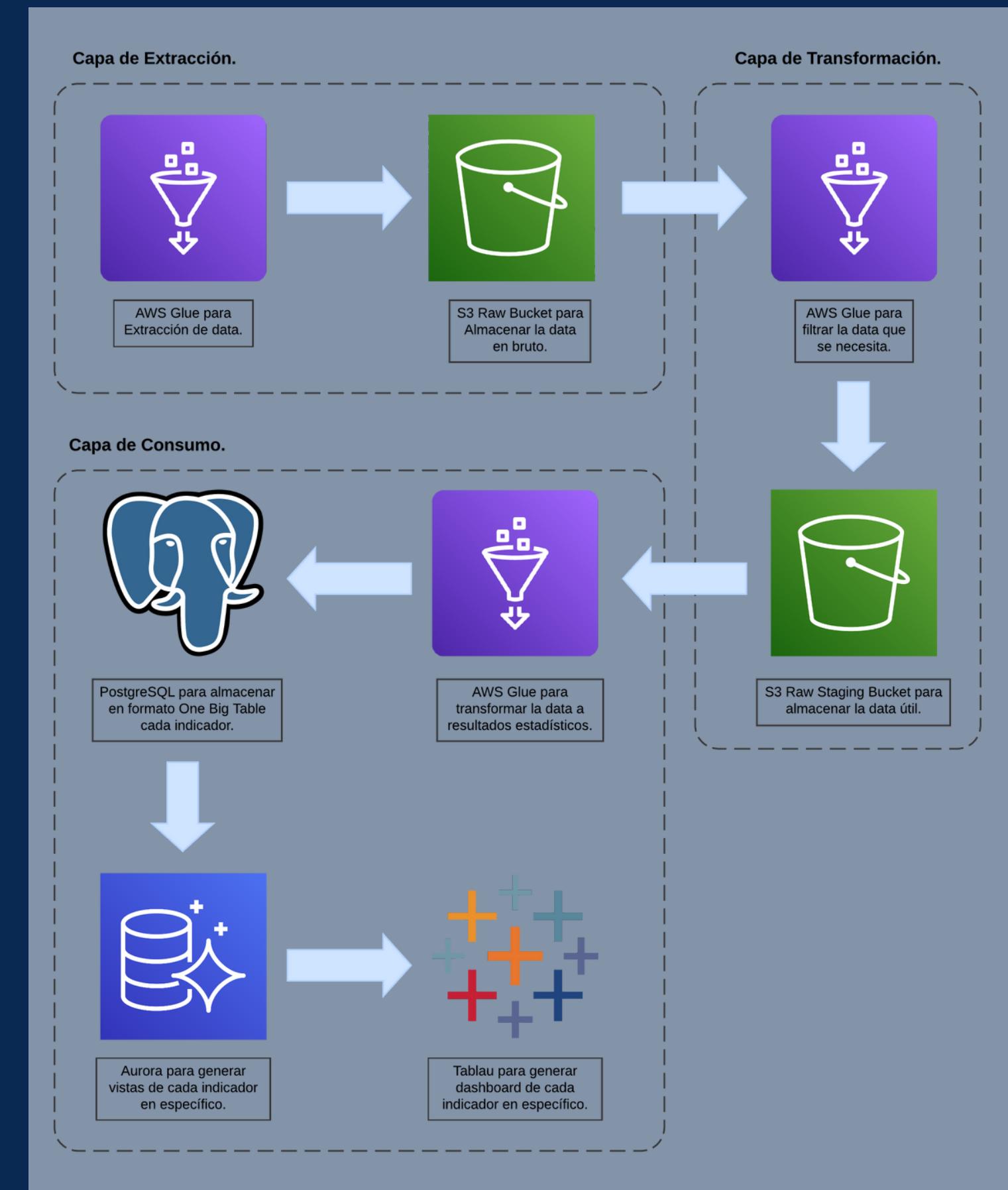
ESTE INFORME DESTACA LA NECESIDAD DE MEJORAR LA GESTIÓN Y PROCESOS ESTADÍSTICOS DE DATOS, PROponiendo SOLUCIONES COMO LA ACTUALIZACIÓN EFICIENTE DE DATOS Y LA AUTOMATIZACIÓN EN SU RECOPILACIÓN. SE ENFATIZA LA IMPORTANCIA DE LA ESTANDARIZACIÓN DE BASES DE DATOS, LA ADOPCIÓN DE UNA NUEVA NOMENCLATURA Y UN DISEÑO DE ENTIDAD-RELACIÓN MEJORADO PARA UNA MAYOR EFICIENCIA OPERATIVA. ESTAS REFORMAS BUSCAN NO SOLO SOLUCIONES INMEDIATAS SINO TAMBIÉN UN DESARROLLO SOSTENIBLE DE LOS SISTEMAS DE DATOS, APUNTANDO HACIA UNA INFRAESTRUCTURA ROBUSTA, ADAPTABLE Y ESCALABLE QUE PERMITA A LA INSTITUCIÓN LIDERAR CON INNOVACIÓN EN UN ENTORNO DE CONSTANTE CAMBIO.

CONCLUSIÓN



- LA FIGURA ILUSTRA UN FLUJO DE TRABAJO A IMPLEMENTAR EN LA UNIDAD DE DATOS, QUE COMIENZA CON LA EXTRACCIÓN DE DATOS Y CULMINA EN VISUALIZACIONES GRÁFICAS PARA LA PLATAFORMA OBSERVA.
- INCLUYE EL USO INTEGRADO DE TECNOLOGÍAS AWS Y SOFTWARE DE TERCEROS, COOPERANDO PARA PRODUCIR INFORMACIÓN PRECISA Y ACTUALIZADA.
- A PARTIR DE LA EXTRACCIÓN, LOS DATOS SE SOMETEN A UN PROCESO DE LIMPIEZA Y ALMACENAMIENTO ANTES DE SU TRANSFORMACIÓN.
- LUEGO, ESTOS DATOS TRANSFORMADOS SE COMBINAN CON OTROS CONJUNTOS PARA GENERAR DASHBOARD.
- ESTE FLUJO ASEGURA QUE LAS REPRESENTACIONES GRÁFICAS FINALES SEAN TANTO CONFIABLES COMO RELEVANTES PARA EL USUARIO FINAL.

ANEXO



PRESENTACION FINAL PRÁCTICA 2024

JOSÉ GUILLERMO SEPÚLVEDA SALAZAR
VIERNES 15 DE MARZO DEL 2024