Indicador de Justa Asignación: Riesgos y Contrataciones Irregulares en Infraestructura

Grupo de trabajo: Alan Yovany Gasca, Christian David González, Jolman Salinas

Proyecto aplicado en analítica de datos

MIAD

Universidad de los Andes

> El presente proyecto busca asistir a las entidades gubernamentales de Colombia en la detección y prevención de prácticas indebidas en la asignación de contratos públicos de infraestructura. A través del análisis de datos provenientes de tres fuentes principales - SECOP Integrado, PACO y datos públicos de transparencia, lucha contra la corrupción, sanciones y multas - el objetivo es minimizar la exposición de estas entidades a contratos poco transparentes, incumplidos o corruptos.

Proceso ETL

Datos de contratación

Para llevar a cabo este análisis, se implementa un proceso ETL (Extracción, Transformación y Carga) que consta de varias etapas:

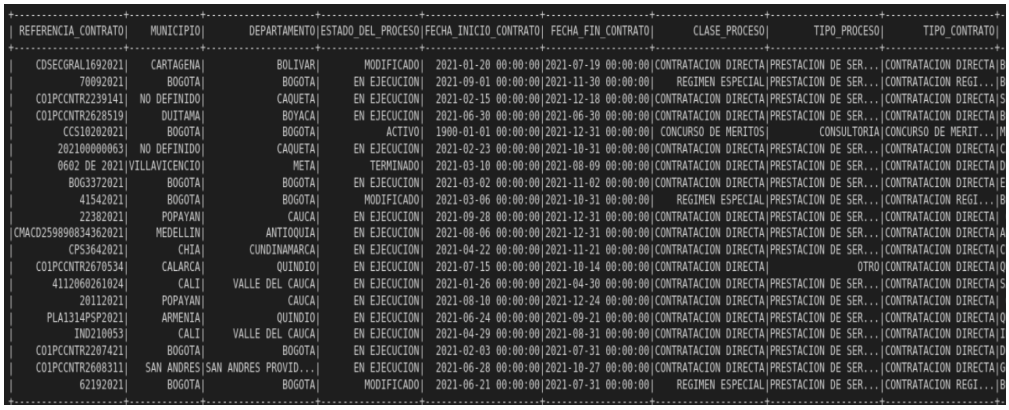
- **Extracción de datos**: La función read_csv_with_pyspark se utiliza para leer archivos CSV almacenados en la carpeta 'data', utilizando el separador '|'. Esta función devuelve un DataFrame de PySpark con la información del archivo CSV.
- **Análisis de calidad de datos**: La función analyze_data_quality se encarga de realizar un análisis básico de la calidad de los datos en el DataFrame proporcionado. Esto incluye el conteo de registros, registros nulos y duplicados, estadísticas descriptivas para columnas numéricas y la identificación de valores atípicos basados en el rango intercuartil.
- **Limpieza de nulos y duplicados**: La función limpiar_nulos_y_duplicados recibe un DataFrame y una lista de columnas. Esta función elimina los registros que contienen valores nulos en las columnas especificadas y también elimina los registros duplicados.
- **Filtrado de categorías de infraestructura**: Se realiza un filtrado de datos en base a las categorías de interés relacionadas con la infraestructura, como "VIVIENDA CIUDAD Y TERRITORIO", "TRANSPORTE", "MINAS Y ENERGIA" y "AMBIENTE Y DESARROLLO SOSTENIBLE".
- **Análisis de calidad de datos en el DataFrame filtrado**: Se vuelve a realizar un análisis de calidad de datos en el DataFrame filtrado para identificar posibles problemas en los datos después de la transformación.
- **Limpieza de nulos y duplicados en el DataFrame filtrado**: Se utiliza la función limpiar_nulos_y_duplicados con las columnas de referencia 'REFERENCIA_CONTRATO' y

'NIT_ENTIDAD' para eliminar registros nulos y duplicados en el DataFrame paco_secop_df. Se guarda el resultado en infraestructura_df_limpio.

- **Filtrado de los últimos años**: Se aplica la función filtrar_ultimos_anos al DataFrame infraestructura_df_limpio para conservar únicamente los registros correspondientes a los últimos 2 años. Se guarda el resultado en infraestructura_df_limpio_2_anos.

- **Selección de columnas relevantes**: Se utiliza la función seleccionar_columnas con una lista de columnas de interés (columnas_a_conservar) para reducir el número de columnas en el DataFrame infraestructura_df_limpio_2_anos. Se guarda el resultado en infraestructura_df_seleccionado.

- **Escritura de resultados en CSV**: Finalmente, el DataFrame resultante, infraestructura_df_seleccionado, se guarda en un archivo CSV llamado "contratos_infraestructura_df.csv" en la carpeta "etl_data".



REFERENCIA_CONTRATO	MUNICIPIO	DEPARTAMENTO	ESTADO_DEL_PROCESO	FECHA_INICIO_CONTRATO	FECHA_FIN_CONTRATO	CLASE_PROCESO	TIPO_PROCESO	TIPO_CONTRATO
CDSECGRAL1692021	CARTAGENA	BOLIVAR	MODIFICADO	2021-01-20 00:00:00	2021-07-19 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA|B
70092021	BOGOTA	BOGOTA	EN EJECUCION	2021-09-01 00:00:00	2021-11-30 00:00:00	REGIMEN ESPECIAL	PRESTACION DE SER...	CONTRATACION REGI...|B
CO1PCCNTR2239141	NO DEFINIDO	CAQUETA	EN EJECUCION	2021-02-15 00:00:00	2021-12-18 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA|S
CO1PCCNTR2628519	DUITAMA	BOYACA	EN EJECUCION	2021-06-30 00:00:00	2021-06-30 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA|B
CCS10202021	BOGOTA	BOGOTA	ACTIVO	1900-01-01 00:00:00	2021-12-31 00:00:00	CONCURSO DE MERITOS	CONSULTORIA	CONCURSO DE MERIT...|M
202100000063	NO DEFINIDO	CAQUETA	EN EJECUCION	2021-02-23 00:00:00	2021-10-31 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA|C
0602 DE 2021	VILLAVICENCIO	META	TERMINADO	2021-03-10 00:00:00	2021-08-09 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA|D
BOG3372021	BOGOTA	BOGOTA	EN EJECUCION	2021-03-02 00:00:00	2021-11-02 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA|E
41542021	BOGOTA	BOGOTA	MODIFICADO	2021-03-06 00:00:00	2021-10-31 00:00:00	REGIMEN ESPECIAL	PRESTACION DE SER...	CONTRATACION REGI...|B
22382021	POPAYAN	CAUCA	EN EJECUCION	2021-09-28 00:00:00	2021-12-31 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA|
CMACD259890834362021	MEDELLIN	ANTIOQUIA	EN EJECUCION	2021-08-06 00:00:00	2021-12-31 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA|A
CPS3642021	CHIA	CUNDINAMARCA	EN EJECUCION	2021-04-22 00:00:00	2021-11-21 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA|C
CO1PCCNTR2670534	CALARCA	QUINDIO	EN EJECUCION	2021-07-15 00:00:00	2021-10-14 00:00:00	CONTRATACION DIRECTA	OTRO	CONTRATACION DIRECTA|O
4112060261024	CALI	VALLE DEL CAUCA	EN EJECUCION	2021-01-26 00:00:00	2021-04-30 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA|S
20112021	POPAYAN	CAUCA	EN EJECUCION	2021-08-10 00:00:00	2021-12-24 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA|
PLA1314PSP2021	ARMENIA	QUINDIO	EN EJECUCION	2021-06-24 00:00:00	2021-09-21 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA|O
IND210053	CALI	VALLE DEL CAUCA	EN EJECUCION	2021-04-29 00:00:00	2021-08-31 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA|I
CO1PCCNTR2207421	BOGOTA	BOGOTA	EN EJECUCION	2021-02-03 00:00:00	2021-07-31 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA|D
CO1PCCNTR2608311	SAN ANDRES	SAN ANDRES PROVID...	EN EJECUCION	2021-06-28 00:00:00	2021-10-27 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA|G
62192021	BOGOTA	BOGOTA	MODIFICADO	2021-06-21 00:00:00	2021-07-31 00:00:00	REGIMEN ESPECIAL	PRESTACION DE SER...	CONTRATACION REGI...|B

Figura 1. Vista del DataFrame de contratos transformado

A través de este proceso ETL, se garantiza la calidad y relevancia de los datos analizados, permitiendo a las entidades gubernamentales colombianas identificar y prevenir situaciones de riesgo en la asignación de contratos públicos de infraestructura. El resultado de este análisis es un archivo CSV que contiene información detallada y depurada sobre los contratos de infraestructura en los últimos dos años, facilitando la toma de decisiones y el monitoreo de posibles irregularidades en el proceso de contratación.

Datos de entidades

En el proceso de ETL segenera tambien un dataframe de entidades que permita perfilarlas para los análisis posteriores, para ello se implementan las siguientes funciones:

- **agregar_por_nit_entidad**: Esta función recibe un DataFrame y realiza agregaciones a nivel de entidad (NIT_ENTIDAD y NOMBRE_ENTIDAD) utilizando diversas métricas, como el número de contratos, la suma y el promedio del valor total de los contratos, el último contrato firmado, la cantidad de departamentos, estados de proceso, clases de proceso, tipos de proceso, familias y clases involucradas. La función también calcula la cantidad de meses transcurridos desde el último contrato.

- **pivotar_por_columna**: Esta función recibe un DataFrame y una columna, y realiza una operación de pivoteo en función de los valores distintos presentes en la columna especificada. El resultado es un DataFrame con una columna por cada valor distinto encontrado, y el conteo de registros por entidad para cada valor.

- **unir_dataframes**: Esta función recibe dos DataFrames y realiza una unión 'inner' entre ellos utilizando las columnas 'NIT_ENTIDAD' y 'NOMBRE_ENTIDAD' como claves de unión.
- **aggregate_multas_data**: Esta función recibe un DataFrame de multas y realiza agregaciones a nivel de entidad (nit_entidad), calculando el número de multas, la suma y el promedio del valor de las sanciones, y los meses transcurridos desde la última multa.
- **left_join_dataframes**: Esta función recibe dos DataFrames y las columnas clave para realizar una unión 'left' entre ellos.

Posteriormente, se aplican estas funciones al DataFrame infraestructura_df_seleccionado para generar un perfil detallado de las entidades involucradas en los contratos de infraestructura. Se realiza el pivoteo y la unión de los DataFrames en función de las columnas 'DEPARTAMENTO', 'ESTADO_DEL_PROCESO', 'CLASE_PROCESO', 'TIPO_PROCESO' y 'NOMBRE_CLASE', obteniendo así un DataFrame agregado y pivotado (infraestructura_df_agregado_y_pivotado) que contiene información clave sobre las entidades y su relación con los contratos de infraestructura.

Finalmente, se realiza una unión 'left' entre el DataFrame de entidades agregado y pivotado y el DataFrame de multas agregado, utilizando la columna 'nit_entidad' como clave de unión. El resultado es un DataFrame que contiene información detallada y depurada sobre las entidades involucradas en contratos de infraestructura, incluyendo información sobre multas y sanciones, lo que facilita la toma de decisiones y el monitoreo de posibles irregularidades en el proceso de contratación.