PYSPARK [EJERCICIO 4]

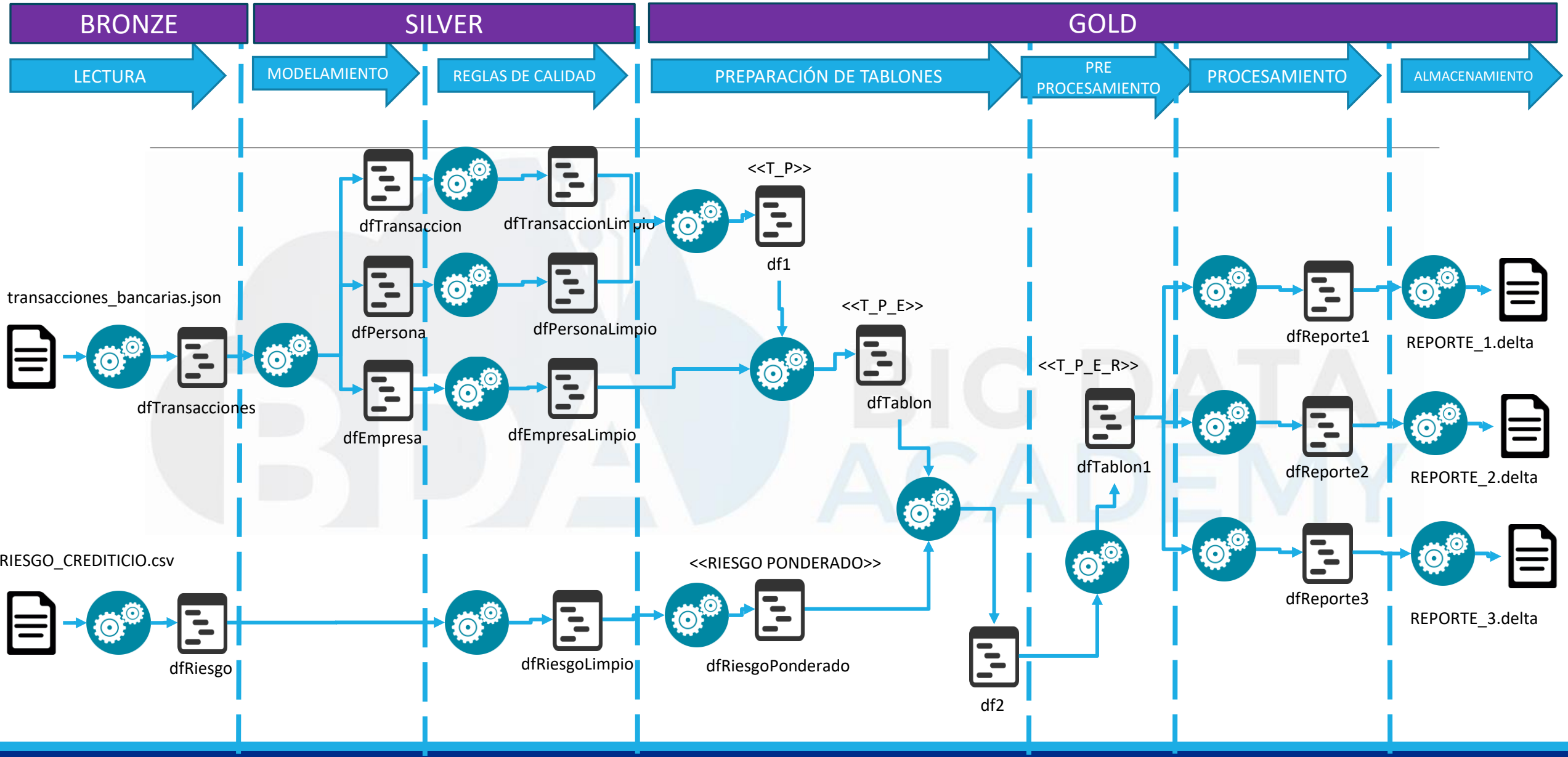
ARQUETIPO AVANZADO DELTA LAKE

Alonso Melgarejo [alonsoraulmgs@gmail.com]

EJERCICIO 4

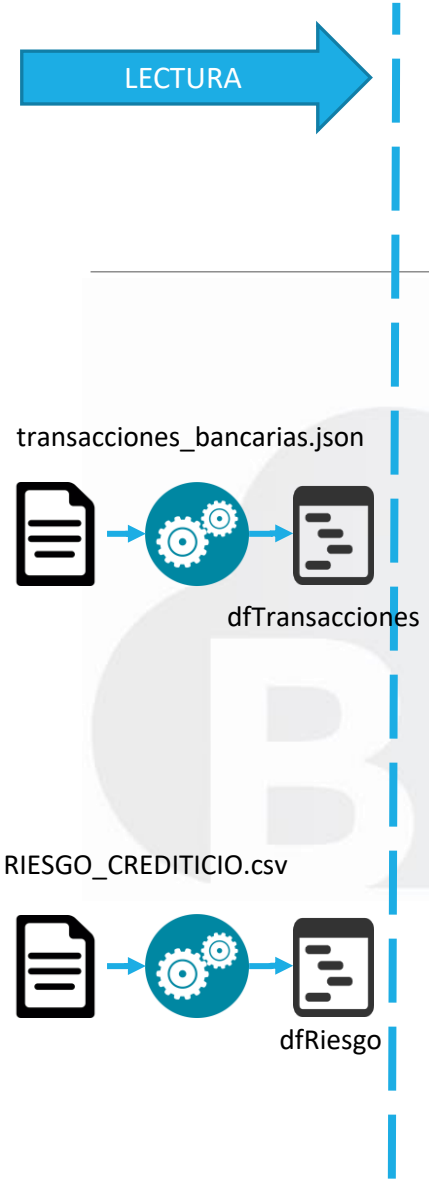
IMPLEMENTAR EL SIGUIENTE PROCESO

Ejercicio 4


BRONZE
SILVER
GOLD
LECTURA
MODELAMIENTO
REGLAS DE CALIDAD
PREPARACIÓN DE TABLONES
PRE PROCESAMIENTO
PROCESAMIENTO
ALMACENAMIENTO
transacciones_bancarias.json
dfTransacciones
dfTransaccion
dfPersona
dfEmpresa
dfTransaccionLimpio
dfPersonaLimpio
dfEmpresaLimpio
<<T_P>>
df1
<<T_P_E>>
dfTablon
RIESGO_CREDITICIO.csv
dfRiesgo
dfRiesgoLimpio
<<RIESGO PONDERADO>>
dfRiesgoPonderado
df2
<<T_P_E_R>>
dfTablon1
dfReporte1
dfReporte2
dfReporte3
REPORTE_1.delta
REPORTE_2.delta
REPORTE_3.delta

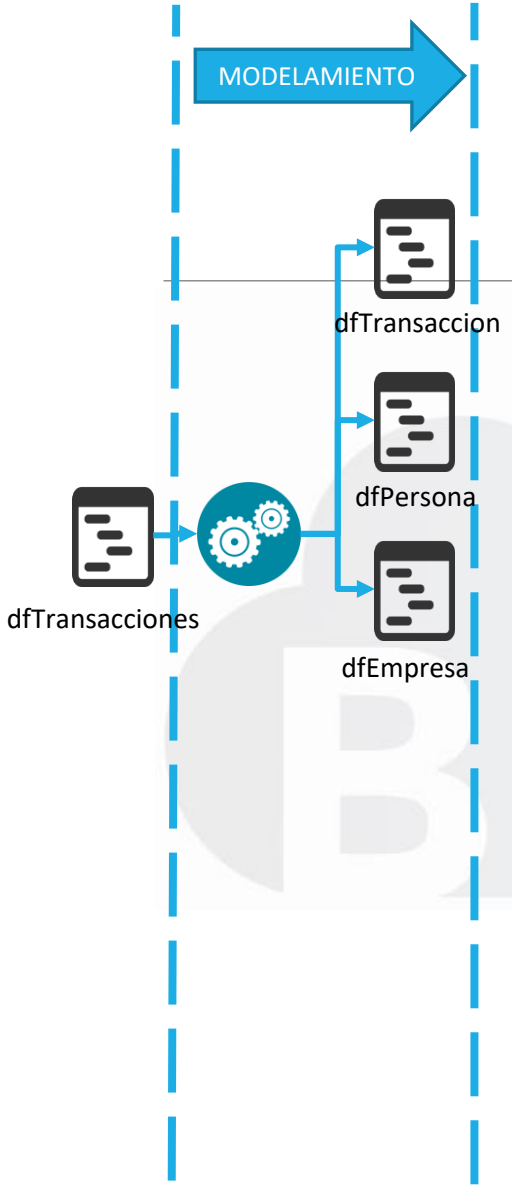
Paso 1: Desplegar la taxonomía Delta Lake

- dbfs://FileStore/_pyspark/**EJERCICIO_4**/
    - bronze
        - riesgo_crediticio
            - RIESGO_CREDITICIO.data
        - transacciones_bancarias
            - transacciones_bancarias.json
    - silver
        - riesgo_crediticio
        - persona
        - empresa
        - transaccion
    - gold
        - REPORTE_1
        - REPORTE_2
        - REPORTE_3



PASO 2: LECTURA

Leer los archivos de datos en DATAFRAMES



PASO 3: MODELAMIENTO

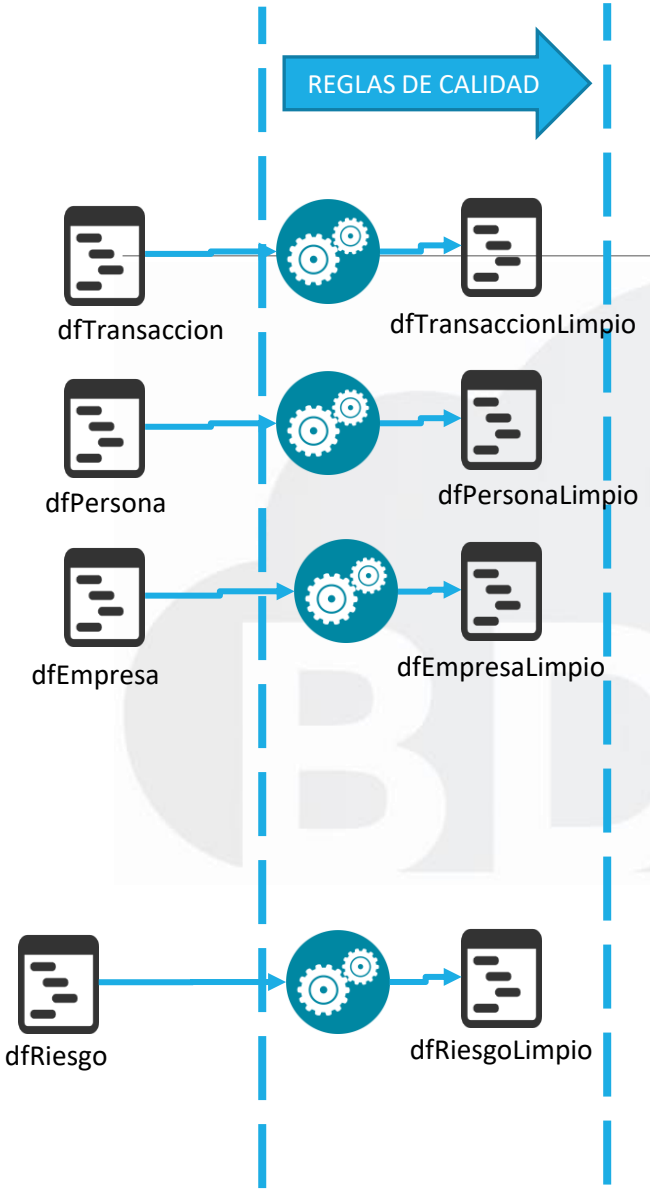
Estructurar el dfTransacciones

dfTransaccion
ID_PERSONA
ID_EMPRESA
MONTO
FECHA

dfPersona
ID_PERSONA
NOMBRE_PERSONA
EDAD
SALARIO

dfEmpresa
ID_EMPRESA
NOMBRE_EMPRESA

PASO 4: REGLAS DE CALIDAD



PERSONA	
ID_PERSONA	NO NULO
SALARIO	RANGO: [0, 100000>
EDAD	RANGO: <0, 60>

RIESGO	
ID_CLIENTE	NO NULO
RIESGO_CENTRAL_1	RANGO: [0, 1]
RIESGO_CENTRAL_2	RANGO: [0, 1]
RIESGO_CENTRAL_3	RANGO: [0, 1]

EMPRESA	
ID_EMPRESA	NO NULO

TRANSACCION	
ID_PERSONA	NO NULO
ID_EMPRESA	NO NULO
MONTO	RANGO: [0, 100000>

PASO 5: CREACIÓN DE UDF

dfRiesgoPonderado
ID_CLIENTE
RIESGO_PONDERADO

FUNCION **calcularRiesgoPonderado**(riesgo1, riesgo2, riesgo3):

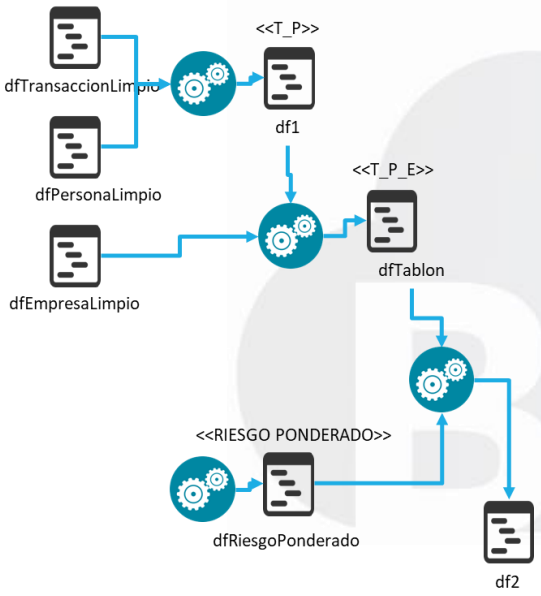
resultado = 0

resultado = (2 x riesgo1 + 3 x riesgo2 + 2 x riesgo3) / 7

RETORNAR resultado

PASO 6: PREPARACIÓN DE TABLONES

PREPARACIÓN DE TABLONES



DATAFRAME: df1

CAMPO	PROVIENE DE
ID_PERSONA	dfTransaccionLimpio
NOMBRE_PERSONA	dfPersonaLimpio
EDAD_PERSONA	dfPersonaLimpio
SALARIO_PERSONA	dfPersonaLimpio
ID_EMPRESA_TRANSACCION	dfTransaccionLimpio
MONTO_TRANSACCION	dfTransaccionLimpio
FECHA_TRANSACCION	dfTransaccionLimpio

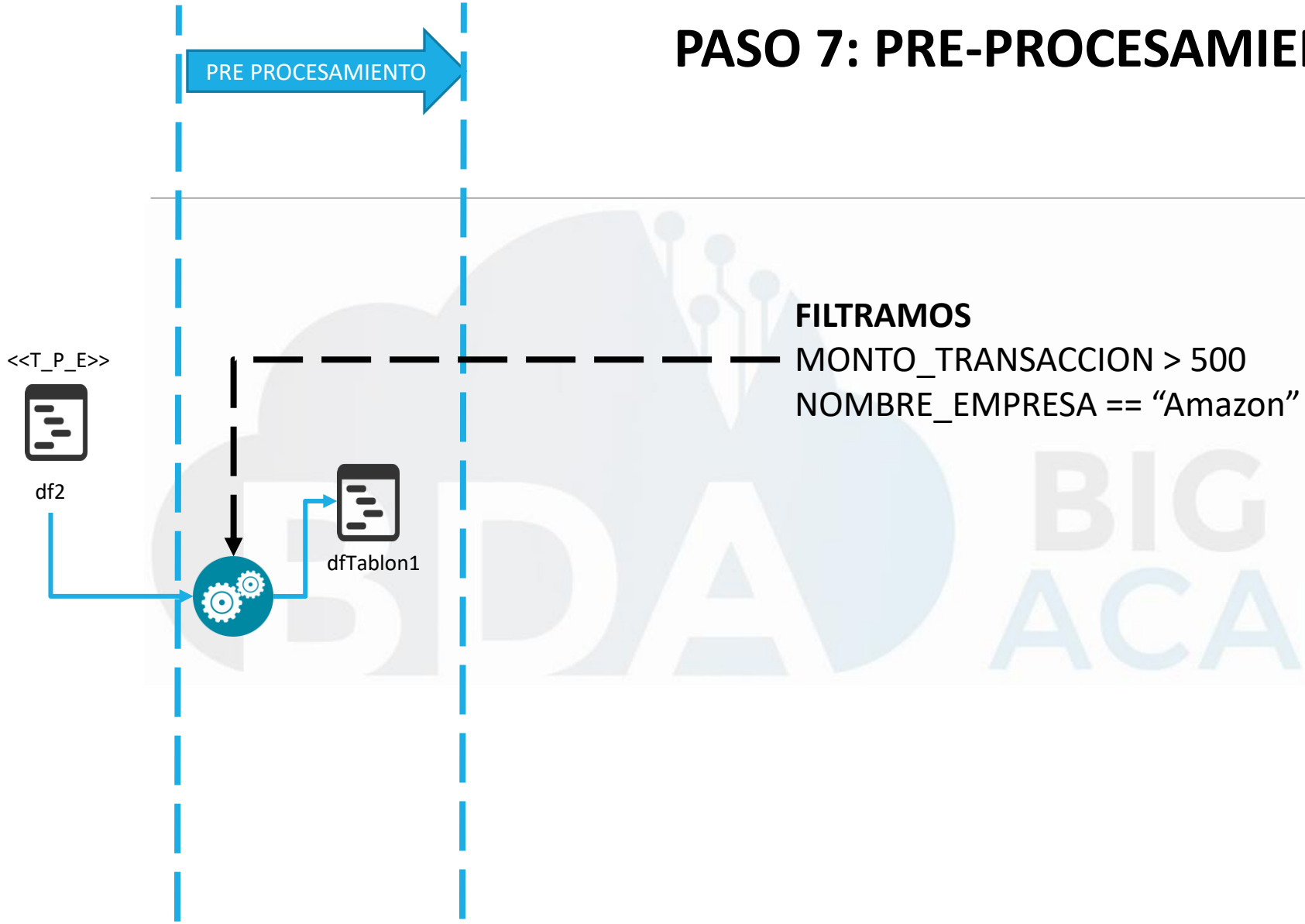
DATAFRAME: dfTablon

CAMPO	PROVIENE DE
ID_PERSONA	df1
NOMBRE_PERSONA	df1
EDAD_PERSONA	df1
SALARIO_PERSONA	df1
ID_EMPRESA_TRANSACCION	df1
NOMBRE_EMPRESA	dfEmpresaLimpio
MONTO_TRANSACCION	df1
FECHA_TRANSACCION	df1

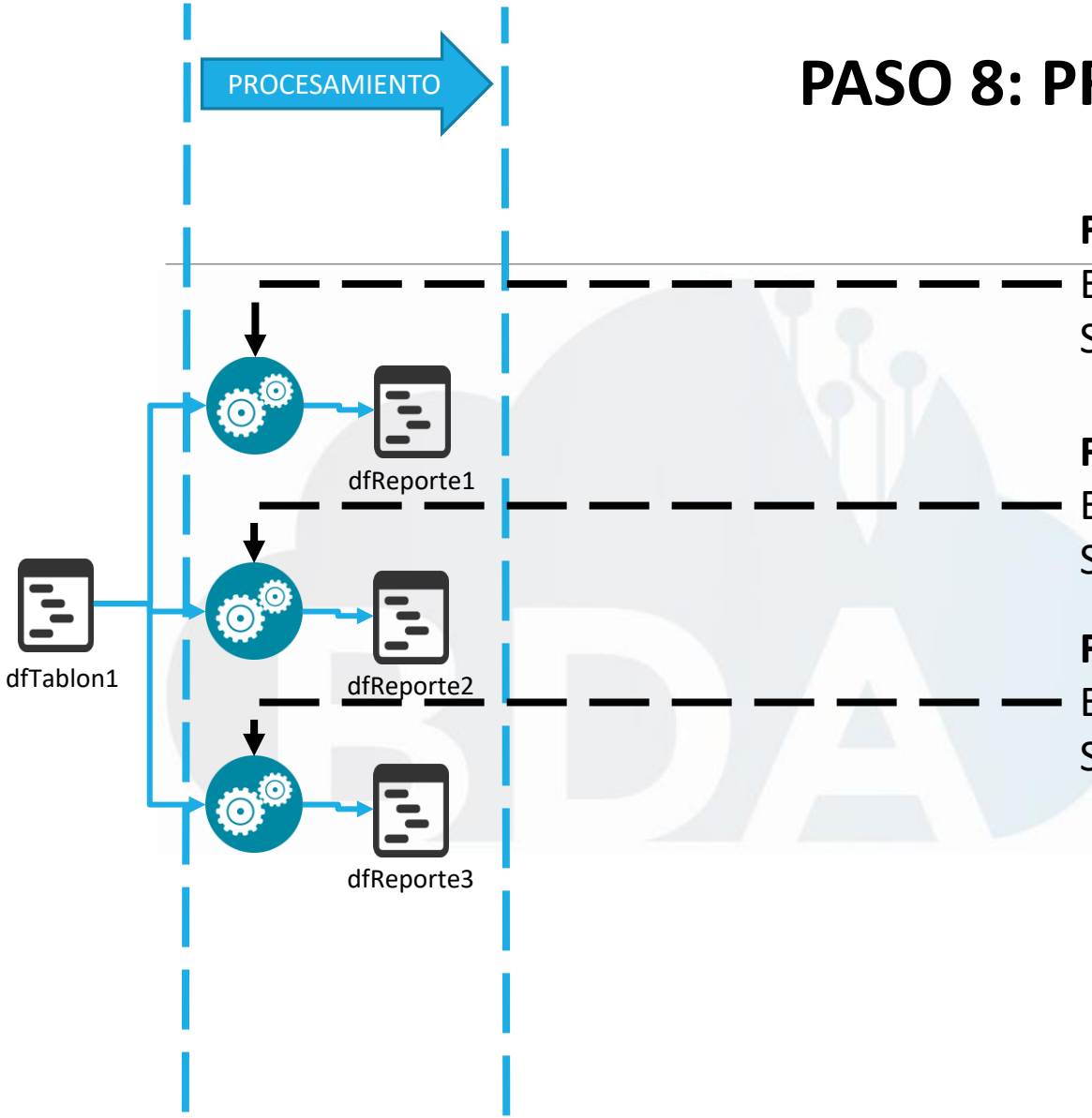
DATAFRAME: df2

CAMPO	PROVIENE DE
ID_PERSONA	dfTablon
NOMBRE_PERSONA	dfTablon
EDAD_PERSONA	dfTablon
SALARIO_PERSONA	dfTablon
RIESGO_PONDERADO	dfRiesgoPonderado
ID_EMPRESA_TRANSACCION	dfTablon
NOMBRE_EMPRESA	dfTablon
MONTO_TRANSACCION	dfTablon
FECHA_TRANSACCION	dfTablon

PASO 7: PRE-PROCESAMIENTO



PASO 8: PROCESAMIENTO

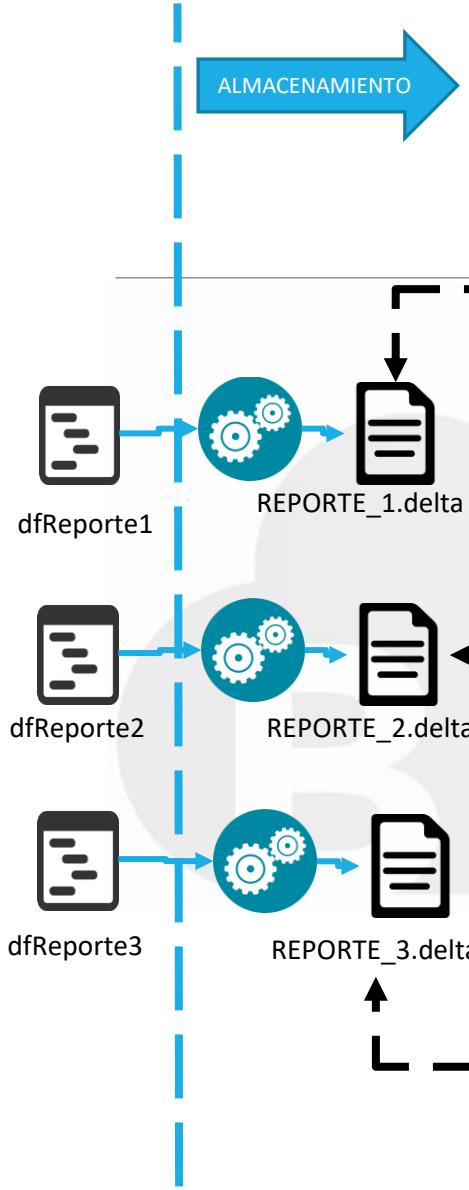


FILTRAMOS
EDAD_PERSONA = [30, 39]
SALARIO_PERSONA = [1000, 5000]

FILTRAMOS
EDAD_PERSONA = [40, 49]
SALARIO_PERSONA = [2500, 7000]

FILTRAMOS
EDAD_PERSONA = [50, 60]
SALARIO_PERSONA = [3500, 10000]

PASO 9: ALMACENAMIENTO



ALMACENAMIENTO
Formato: **delta**
Directorio: dbfs:///FileStore/_pyspark/EJERCICIO_4/gold/REPORTE_1

ALMACENAMIENTO
Formato: **delta**
Directorio: dbfs:///FileStore/_pyspark/EJERCICIO_4/gold/REPORTE_2

ALMACENAMIENTO
Formato: **delta**
Directorio: dbfs:///FileStore/_pyspark/EJERCICIO_4/gold/REPORTE_3