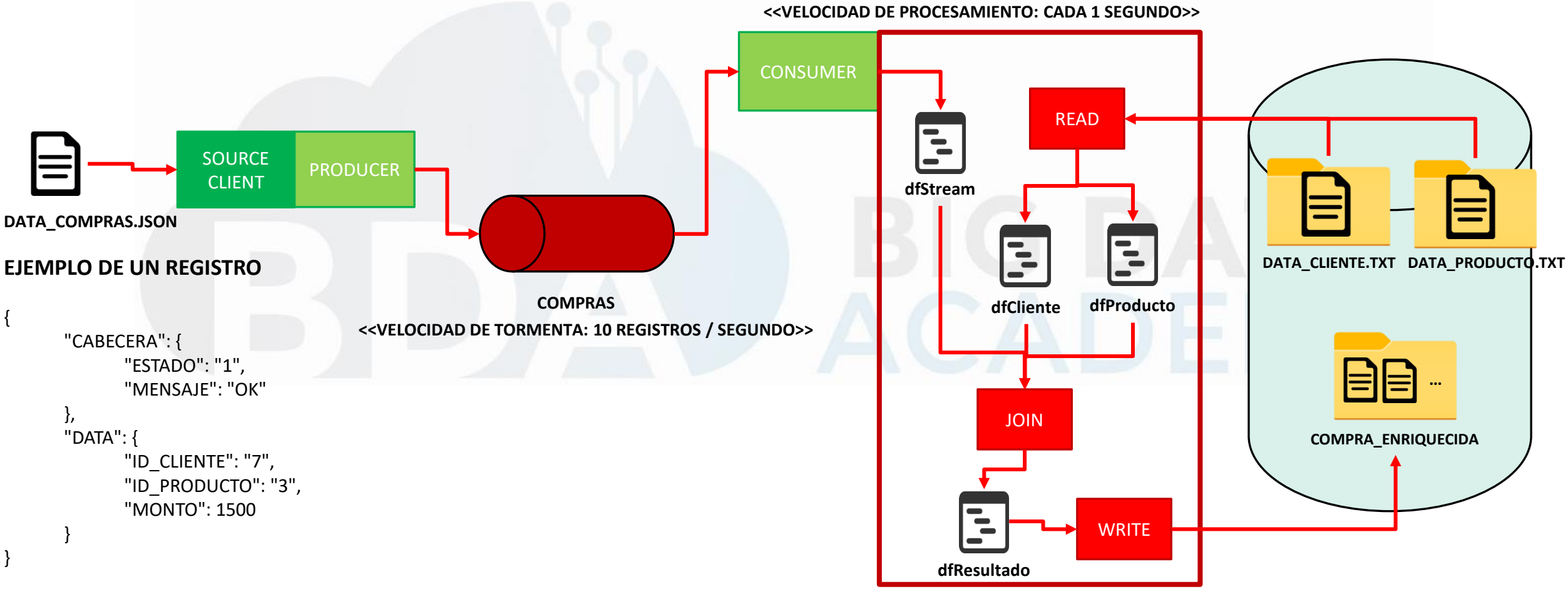
EJERCICIO 6

PROCESAMIENTO REAL TIME

Implementar el siguiente proceso



EJEMPLO DE UN REGISTRO

```
{
        "CABECERA": {
                "ESTADO": "1",
                "MENSAJE": "OK"
        },
        "DATA": {
                "ID_CLIENTE": "7",
                "ID_PRODUCTO": "3",
                "MONTO": 1500
        }
}
```

PASO 1: INSTALAR EL CLÚSTER KAFKA Y CREAR EL TÓPICO

1. NOTEBOOK 1: Instalar en el CLÚSTER las librerías MAVEN de lectura y procesamiento de KAFKA, descargar KAFKA, instalarlo e iniciar ZOOKEEPER.
2. NOTEBOOK 2: Iniciar KAFKA.
3. NOTEBOOK 3: Crear el tópico “compras” e iniciar un CONSUMER de consola.
4. NOTEBOOK 4: Crear un PRODUCER de consola, enviar el mensaje “Hola mundo” y verificar en el NOTEBOOK 3 que el mensaje haya llegado al CONSUMER de consola.

PASO 2: IMPLEMENTAR EL CLIENT SOURCE Y EL PRODUCER

En un NOTEBOOK 5:

- Implementar el CLIENT SOURCE que emule la tormenta de datos del archivo “DATA_COMPRAS.json”, con una velocidad de 10 registros por segundo.
- Implementar el PRODUCER que escriba las transacciones en el tópico “compras”.

PASO 3: IMPLEMENTAR EL CONSUMER

En un NOTEBOOK 6:

- Leer el archivo “DATA_CLIENTE” en el dataframe “dfCliente”
- Leer el archivo “DATA_PRODUCTO” en el dataframe “dfProducto”
- Implementar el CONSUMER que procese los datos cada 1 segundo y los coloque dentro del dataframe estructurado “dfStream” con los campos “ID_CLIENTE | ID_PRODUCTO | MONTO”

PASO 4: IMPLEMENTAR EL PROCESO

En el mismo NOTEBOOK 6:

- Construir el dataframe resultante “dfResultado” que tenga la información enriquecida de las compras: “ID_CLIENTE | NOMBRE_CLIENTE | EDAD_CLIENTE | ID_PRODUCTO | NOMBRE_PRODUCTO | MONTO”
- Almacenar el dataframe “dfResultado” en formato parquet en la ruta “dbfs:///FileStore/_spark/output/COMPRA_ENRIQUECIDA”