Ejercicio Big Data Aplicado – 2ª Evaluación

Dispones de un archivo de texto (firewall_log.txt) que contiene registros de un firewall basado en IPTables. Cada línea del log es una cadena de texto de longitud variable y con un número indeterminado de campos. Los registros pueden incluir información de tráfico, administración o configuraciones.

Objetivo:
Detectar las IPs que han intentado conectar a **múltiples puertos únicos** en un corto período de tiempo, lo que podría indicar un escaneo de puertos. Para ello, deberás:

1. **Leer** el archivo de logs usando RDDs.
2. **Filtrar y limpiar** las líneas que corresponden a eventos de tráfico (por ejemplo, aquellas que contienen la palabra "IPTABLES:").
3. **Parsear** cada línea para extraer al menos la siguiente información:
    - **Timestamp:** La fecha y hora del registro (los primeros tokens de la línea).
    - **Fuente (SRC):** La dirección IP origen.
    - **Puerto destino (DPT):** El puerto al que se intenta conectar.
4. **Transformar** los datos usando:
    - **flatMap:** Para manejar la posibilidad de que una línea pueda generar cero o más registros estructurados (en caso de que contenga más de un evento o múltiples valores en algunos campos).
    - **map:** Para convertir cada registro en un par clave/valor, donde la clave sea la IP de origen y el valor sea el puerto destino y la hora.
    - **reduce (reduceByKey):** Para agregar (por ejemplo, unificar conjuntos) todos los puertos destino únicos asociados a cada IP.
5. **Filtrar y ordenar** los resultados para identificar aquellas IPs que hayan intentado conectarse a más de un cierto número de puertos en una hora (por ejemplo, más de 5 puertos únicos).
6. **Convertir** el RDD final a un DataFrame para facilitar su visualización o posterior almacenamiento.