EXAMEN DATA ENGINEER

Crear un script en Python que se ejecute en Spark 2.4.x que cumpla con las siguientes características:

1. Realizar lectura del csv, adjunto en el correo, para que reconozca la columna: data.
2. Navegación sobre los datos de la columna data de tal forma que se obtengan cualquier metadato en específico, ya que el contenido de la columna son textos en formato JSON.
3. Realizar un agrupamiento por el metadato srcIP y obtener las siguientes métricas: Promedio y desviación estándar del metadato srcBytes, totBytes y length.
4. Realizar un agrupamiento por el metadato: protocol para obtener las métricas de los mismos metadatos del punto anterior. En cuanto al protocolo los valores están jerarquizados por dos niveles, el primer nivel está definido por tcp y udp, las métricas se deben sacar por el segundo nivel. Descubrir los valores del segundo nivel.

Se requiere los siguientes entregables:

1. Código fuente el cual puede ser en una notebook o un archivo simple con formato py
2. Documentación donde se explique la solución. Nada complejo ni respetando algún formato.