Programación Avanzada con Dask

Descripción de la Tarea:

En este laboratorio se deberán utilizar las tres estructuras de datos principales de Dask: Bags, DataFrames y Arrays, para realizar tareas de procesamiento de datos. Deben demostrar su capacidad para trabajar con estas estructuras y comprender cuándo y cómo utilizar cada una de ellas.

Instrucciones: Utiliza Jupyter Notebook para realizar las siguientes actividades:

Actividad 1: Trabajando con Bags

1. Crea un Dask Bag a partir de una lista de diccionarios.
2. Realiza una serie de transformaciones en el Bag, como mapeo, filtrado y agregación.
3. Muestra el resultado final y explora su estructura.

Actividad 2: Manipulación de DataFrames con Dask

1. Carga un conjunto de datos (**a elección**) en formato CSV en un Dask DataFrame.
2. Realiza operaciones de manipulación de datos, como filtrado, agrupamiento y unión de DataFrames.
3. Calcula estadísticas descriptivas en columnas numéricas.

Actividad 3: Trabajo con Arrays en Dask

1. Crea un Dask Array a partir de un conjunto de datos numéricos.
2. Realiza operaciones matriciales y estadísticas en el Array, como multiplicación, suma y cálculo de desviación estándar.
3. Visualiza el Array y los resultados de las operaciones.

Actividad 4: Integración de Estructuras de Datos de Dask

1. Combina el uso de Bags, DataFrames y Arrays de Dask en una tarea que demuestre cómo estas estructuras pueden trabajar juntas.
2. Realiza una operación compleja que involucre al menos dos de las estructuras (por ejemplo, cargar datos en un DataFrame, procesarlos con un Bag y luego realizar un cálculo en un Array).

Actividad 5: Trabajo con Chunks en Dask

1. Carga un conjunto de datos grande (por ejemplo, un archivo CSV grande) en un Dask DataFrame y explora su estructura y tamaño.
2. Divide el DataFrame en trozos (chunks) más pequeños utilizando el método `repartition()`. Establece un tamaño de chunk adecuado para el conjunto de datos.
3. Realiza una operación que involucre la agregación o transformación de los datos en los chunks. Por ejemplo, puedes calcular la suma total de una columna en cada chunk y luego combinar los resultados.
4. Mide y compara el tiempo de ejecución de la operación en los chunks con el tiempo de ejecución si se realizara en el DataFrame completo.
5. Compara el uso de memoria entre el procesamiento por chunks y el procesamiento del DataFrame completo.

Entrega: Se debe presentar un informe en formato Jupyter Notebook que incluya el código, comentarios explicativos y resultados de cada actividad. Deben proporcionar gráficos, visualizaciones y conclusiones cuando sea necesario.