Defensa del proyecto

Teresa Álvarez de Portugal

June 2023

1 Introduction

Aqui responderé a las preguntas relacionadas con la defensa de mi trabajo final de la asignatura de programación paralela y distribuida.

1.1 Preguntas

-¿Cómo se aplicó la programación paralela usando Dask para realizar la carga del conjunto de datos, la eliminación de registros duplicados, y los cálculos de media y desviación estándar?

Lo primero que hice fue importar la biblioteca Dask junto con dask dataframe para poder trabajar. Lo primero que realicé fue leer el csv usando la funcion $dask.dataframe.read_csv()$y guardandolo en un dataframe. Para las otras aplicaciones solo tuve que ir usando las otras funciones, como $.dropna()$ o $.mean()$ o $.std()$. Para que los resultados se viesen de forma adecuada usaba $.compute()$.

-Si tuvieras que escalar este proyecto a un conjunto de datos más grande que no cabría en la memoria de una sola máquina, ¿cómo distribuirías los datos y el trabajo entre diferentes máquinas usando Dask?

Para realizar esto, lo más conveniente sería usar la capacidad de Dask y distribuir tanto los datos como el trabajo entre múltiples máquinas en un clúster. Primero habría que configurar dicho clúster de Dask para realizar lo previamente dicho. Después se dividirían los datos en fragmentos para que se procesen por cada máquina del clúster. Se cargarían los datos distribuidos, se distribuirían las operaciones para aplicar funciones a cada fragmento de los datos y finalmente se ejecutaría el cálculo mediante la función $compute()$.

-Cuando calculaste la matriz de correlación y aplicaste el algoritmo de tu elección, ¿cómo se benefició tu análisis de la programación paralela y distribuida?

En general el análisis se benefició porque al aplicar la matriz de correlación y el algoritmo se obtuvieron resultados más rápidos ya que es mas fácil trabajar con cojuntos grandes lo que me permitió realizar análisis más avanzados