

**Santiago Guerrero Chamorro**

**Seminario De Big Data**

**ELIAS BUITRAGO BOLIVAR**

**Universidad ECCI**

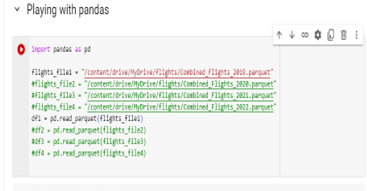
**Facultad de ingeniera**

**2024**

**Tablas Comparativas**

**Pandas**

Con un único archivo, se observó que Pandas no completó correctamente el proceso de importación debido a las limitaciones de recursos:

****

**Polars**

Se cargaron los cuatro archivos con éxito utilizando la librería Polars, como se muestra en la siguiente imagen:



La concatenación y agrupación de los dataframes tomaron un tiempo de procesamiento de 1 minuto:

****

**PySpark**

El cargue de los cuatro archivos se realizó con éxito usando la librería PySpark. El mayor tiempo de procesamiento ocurrió durante la agrupación de los data sets y en las columnas, tardando aproximadamente minuto y medio:

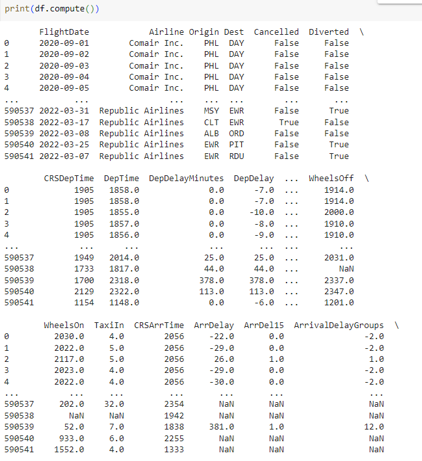


**Dask**

Al cargar archivos con la librería Dask, el proceso fue ágil, como se observa en la siguiente imagen:



El procesamiento para visualizar los datos tomó alrededor de 32 segundos:

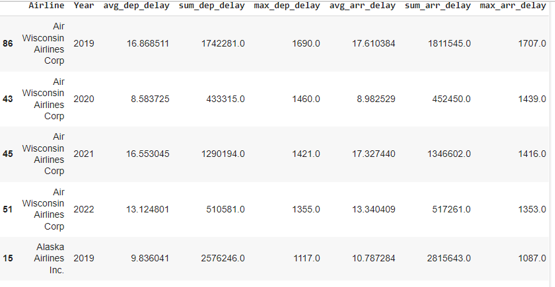


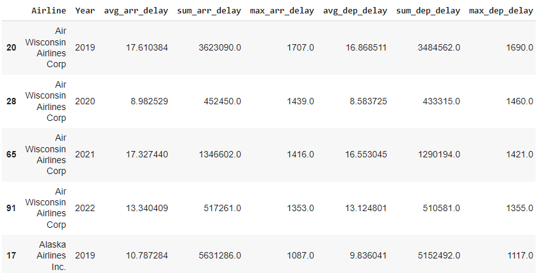
La agrupación de los dataframes mostró un mejor rendimiento, con un tiempo de procesamiento de 1 segundo:

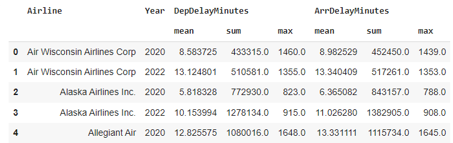


**Análisis de Resultados**

El análisis de resultados con Pandas no fue exitoso debido a que la importación de archivos no se completó, restringida por los recursos del servidor de Google Collaboratory. Por otro lado, las demás librerías, como Polars, pudieron interactuar correctamente, como se evidencia a continuación:







**Conclusiones**

Después de realizar pruebas en Google Collaboratory con las bibliotecas Pandas, Polars, PySpark y Dask, se encontró que todas funcionaron bien dentro de los recursos limitados proporcionados. Sin embargo, Pandas experimentó dificultades al intentar exportar grandes conjuntos de datos en un solo archivo.

En términos de velocidad y capacidad para manejar tareas complejas de análisis de datos, Polars, PySpark y Dask demostraron ser sólidas y consistentes bajo condiciones de recursos restringidos como los que nos brindaba Google Collaboratory. Estas bibliotecas mostraron un rendimiento eficiente y fueron capaces de manejar tamaños de datos crecientes sin comprometer significativamente los tiempos de ejecución.

Por otro lado, Pandas, a pesar de ser ampliamente utilizada y potente, reveló limitaciones específicas al enfrentarse a la exportación de grandes archivos de datos en el entorno de Google Collaboratory. Este hallazgo destaca la importancia de considerar las capacidades de procesamiento y los recursos disponibles al seleccionar la herramienta adecuada para proyectos de análisis de datos en entornos con recursos limitados.

En resumen, mientras que Polars, PySpark y Dask se destacaron por su eficiencia y capacidad para manejar grandes volúmenes de datos en Google Collaboratory, Pandas mostró desafíos en escenarios que requerían la manipulación de archivos de datos extensos. Estos resultados enfatizan la necesidad de evaluar cuidadosamente las herramientas según las necesidades específicas y las limitaciones del entorno de trabajo.