Dask DataFrame Library

By: Mariam Ahmed Towfik

What is Dask DataFrame?

- Collection of pandas dataframes that are working parallelly to work on very large data or many data files on one computer or on distributed computers.

Difference between Pandas DataFrame & Dask's:

Pandas	Dask
Performs good on small data	Performs good on big data or many files
Works in series	Works in parallel

Note: Dask Data frame is a lazy library it only points to the data and doesn't load it in memory until told using :

- *.compute()* ->compute the final result and change its data type to its respective type Dataframe, numpy.
- *.persist()* -> it computes the result while keeping the data type Dask Dataframe, but cache the data so computations will be faster.

Dask methods & attributes:

Import dask.dataframe as dd

Creating Dask Dataframes:

- *dd.read_csv(filepath)* -> to read from csv file
  - **filepath**: the csv files(s) you want to read from.

*It provides loading from other file formats like: parquet, json..etc.

- *dd.from_pandas(data)* -> change pandas dataframe to dask's.

    - **data***:* the dataframe to change.

Writing Dataframe to files:

- *ddf.to_csv(filename) ->* one filename will be created for every partition.
    - **Filename**: the name of the file(s) to write to.

*It provides writing to other file types like parquet.

- *dd.visualize() ->* to visualize the done computation.
- *ddf.map_partitions(func) ->* apply function to each dataframe partition.
    - **func:** the function applied to each partition.
- *ddf.partitions[indices] ->* used to slice the dataframe by partitions.
    - **Indices:** the indices to slice as : [0], [:3]
- *ddf.npartitions ->* returns the number of partitions in the dataframe.
- *ddf._meta ->* returns a dataframe that represents the structure of the dask dataframe(column names, data types). To inspect the data schema.
- *ddf.repartition(npartitions (optional)) ->* repartion dataframe along new division (indices).
    - **npartitions:** number of partitions of output dataframe.

Most Methods that work in pandas work in Dask Like groupby , set_index(), Mathematical computations like .mean..etc.