DA5402: Assignment 10

Let's try to setup a Spark cluster in our laptops using dockerization. We have a Github repo that illustrates the usage of docker images to setup a virtual cluster.

Task 1

The task is to setup the Spark cluster by following the instructions provided on the README page. We will run the example PySpark job which filters a CSV file for invalid entries and pushes the cleansed records into a PGSQL database. Additionally, we will make configuration changes (listed below) to the docker-compose file to witness the changes on the cluster environment.

- Change the number of workers
- Change the memory allocated to the workers
- Change the CPUs allocated to the workers.

Task 2

Run the following toy problems to study to parallelization and distributed computing abilities. We will use an NumPy array of an arbitrary size. We will change the parameters of the array and the task to study the abilities.

- Size of the array
- Number of slices
- Number of workers
- CPU and memory allocated to a worker

We will collect the time taken for completing the task against different choices slices, workers, CPU & memory per worker. The following are the tasks:

- Vector dot product
- Scaling a vector
- Vector addition

NOTE: The docker image build will take some time as it involves multiple downloads. Please build the docker image before coming to the class. The instruction is on the Github README.