Few Words

I will not consider this as a book; instead, a references materials. I have been working with spark for some time and have done enough production implementation. Over time I did face numerous challenges, from basic spark framework to memory optimization. Like everyone else, I also Googled the problem and eventually found the solution. I have learned spark primarily by reading books, researching different forums and youtube videos. I realized I could have avoided most of the problems and utilize my time more effectively if I had some quick references materials.

So I have started gathering all the documents together. So this is the outcome of that effort.

I have kept all the original link on the All Links Snippet page.

I have used two different clusters for all the learning and testings. The first cluster is my local cluster setup using multiple Intel NUC computers using Ubuntu. For more robust usages, I have used ten nodes cluster in GCP. Oh, Man!! The cloud is costly. To minimize the cost, I only use the GCP cluster for more intensive load testing.

To keep the learning more performance-based, I have always used large data set for testing, such as Wiki pages, Yellow taxi, etc. You will get all the details in the '*Get Test Data*' section. I have also generated my test data (RDBMS style). You will get that details too in this section.

If you would like to set up a local spark cluster, follow the '*Set up cluster*' section.

I have tried to avoid all spelling and grammatical mistake. However, if you do find one, please let me know.

I have tested all the codes in my local and GCP cluster. If for any reason the code does not work, then please do drop me a mail - ***contact@sandipan.tech***

Get all the codes from gitHub- ***https://github.com/ghoshm21/spark_book***