Fifa \_19 dataset:

**question** : To store the FIFA19 dataset to enable frequent query and efficient

analysis and

that the dataset size as well as the frequency of load may increase in the future as FIFA’s daily delta datasets are available .

**Solution** : As part of implementing a robust and efficient solution , we can store the data in a hive dynamically partioned table ( on load date) and and load the data into hive partions using spark dataframe everyday. As the data gets stored in a partioned table, quering and processing will be efficient.

While loading the data to a dataframe in the spark job, we can repartion the dataframe with the same number of executers we provide in the config file.we can increase the number of executor as per the requirement when the data gets increased. So that spark

distributes the data efficiently over the cluster and it wont impact the performance.

->We can can convert sort merge join to broadcast join , as we can save the sorting of both the join sides, and read shuffle files locally to save network traffic(if **spark.sql.adaptive.localShuffleReader.enabled is true**) .

I am unable to store the dataset in postgres database using docker because I don’t have docker in my system and I am unable to install as it was not allowing me to do.