**Spark performance tuning handling skew data**

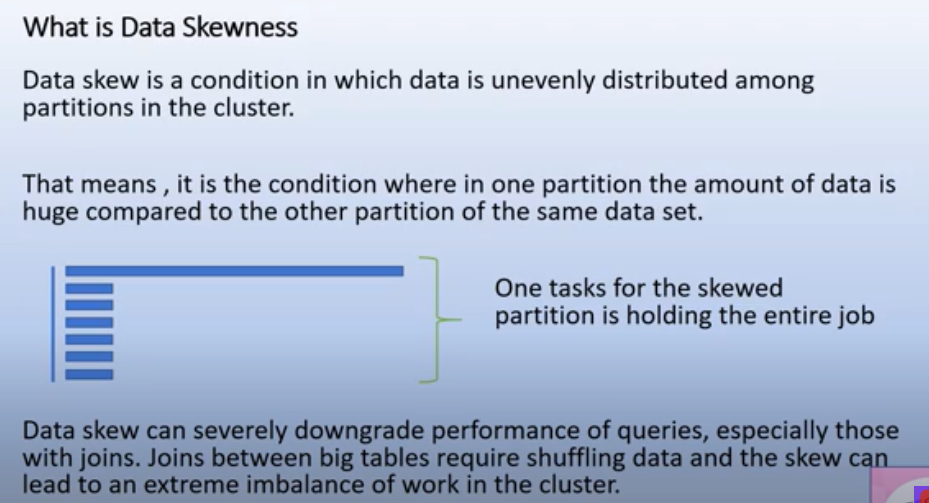
Different techniques available to handle data skew problem

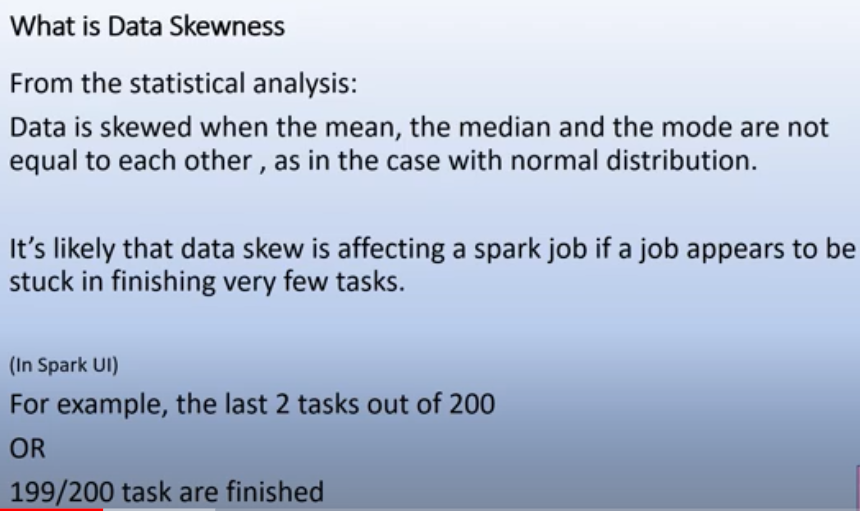
Data in unevenly distributed among the diff partition across the cluster. That means the amount of data in one partition is huge as compare to the other partition of the same data set.

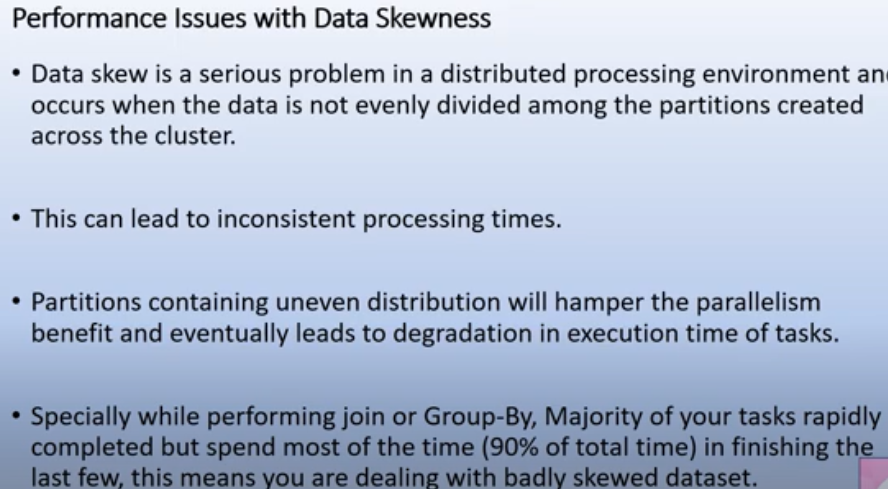
So there is one dataset when you put for computation in spark world and the way partition is created the data skewness is occur where in one partition the amount of data s really high compare to the other partitions

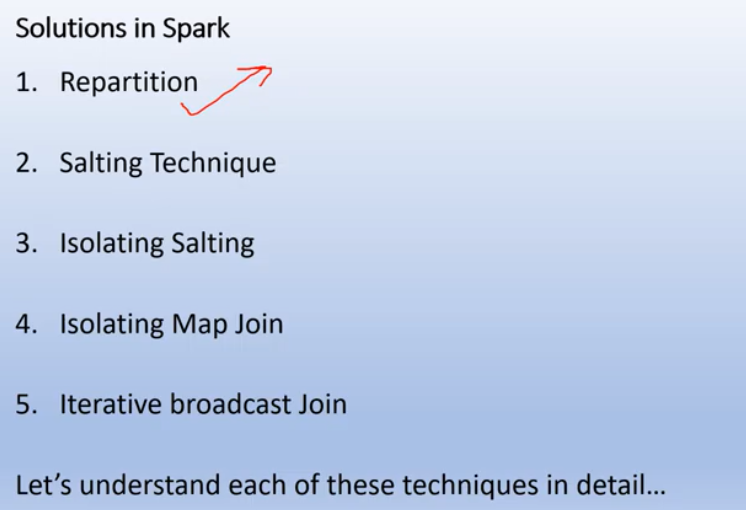
Look at the below situation where spark job is running with diff task with diff partition , all the task with the small partition are finish at the time however they are waiting for the task of the data skewed partition. This skewd data partition hold the entire spark job and you will not able to leverage the right benefits of parallelism in spark. Here parallelism become useless.

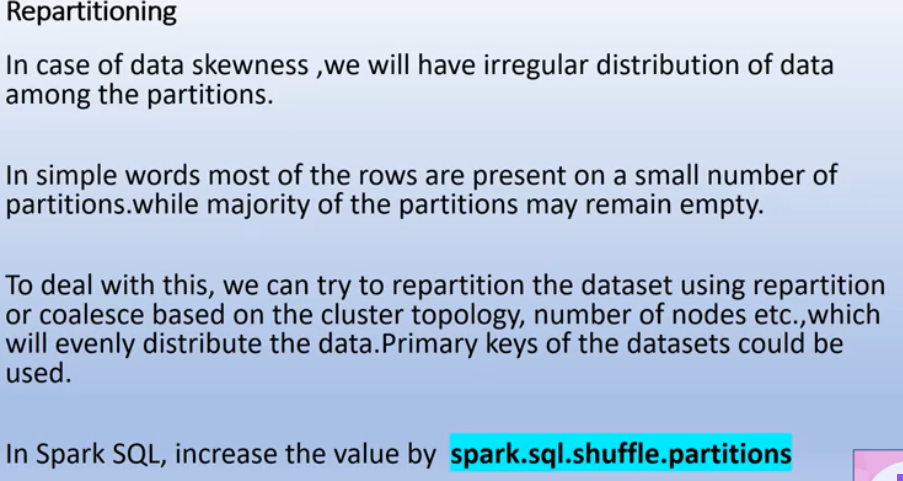
As data skewness downgrade the performance of queries and operation like join, aggregation where shuffling of data is required

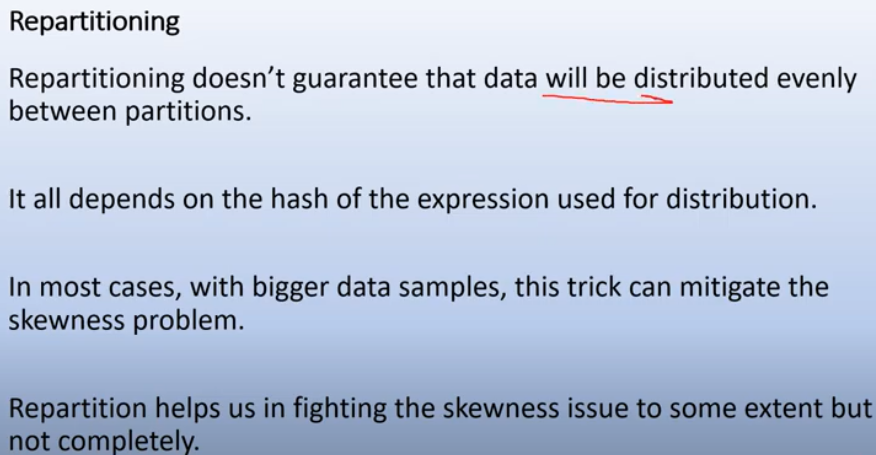




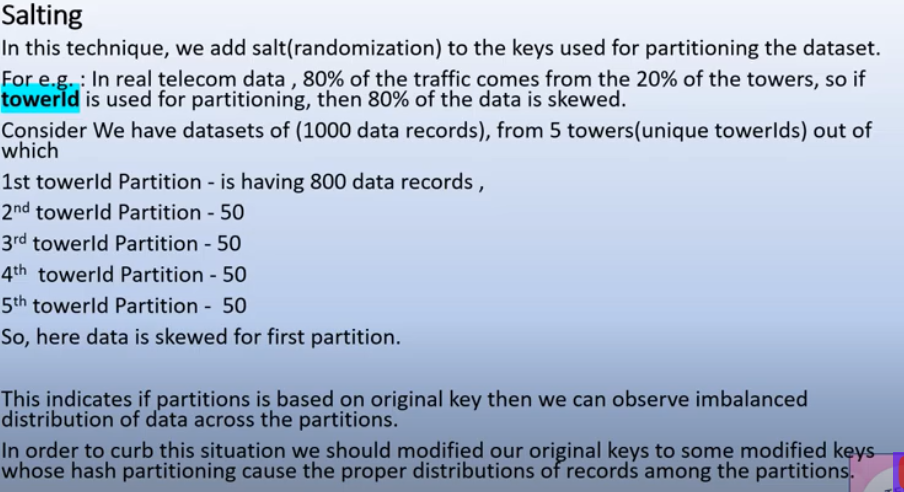








In this technique, we are try to add random value to the usual partition key in the particular dataset so that with the salted keys we can redsitribute the data during the operation like join, aggregation



In this tower usecase, there is skew problem at lasrge extent so to avoid this we have to add some salt(which is just a integer) with the partition key.

