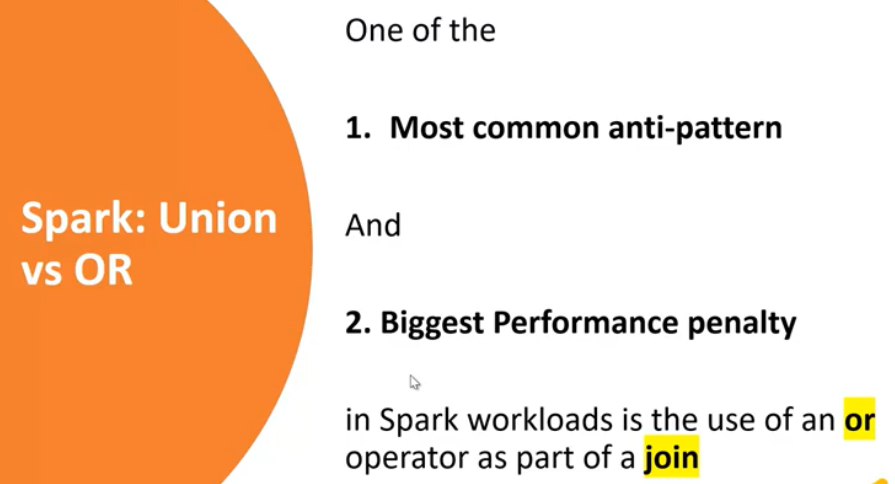
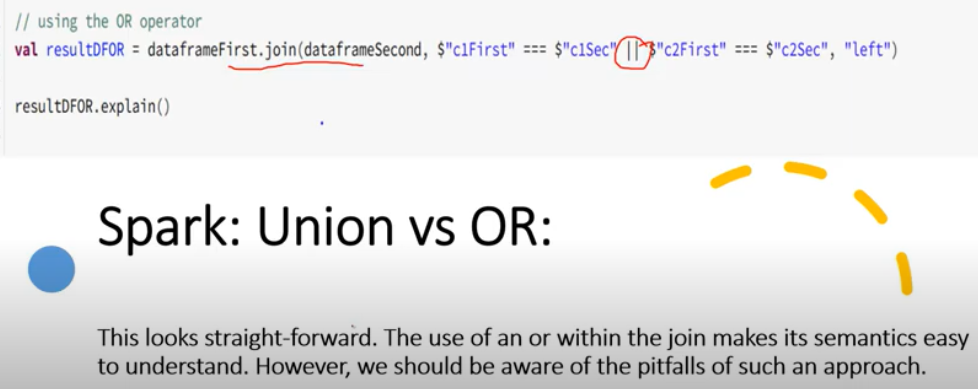
**Spark join performance (union v/s or)**



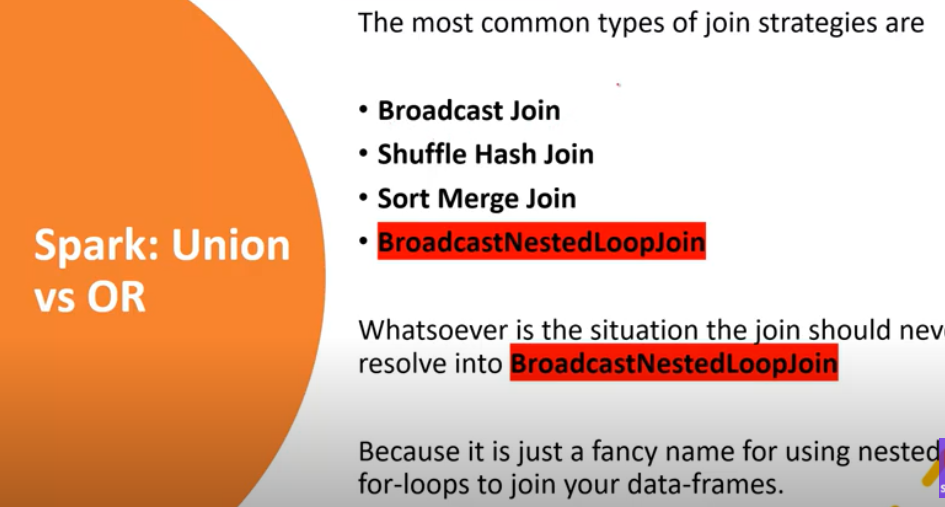


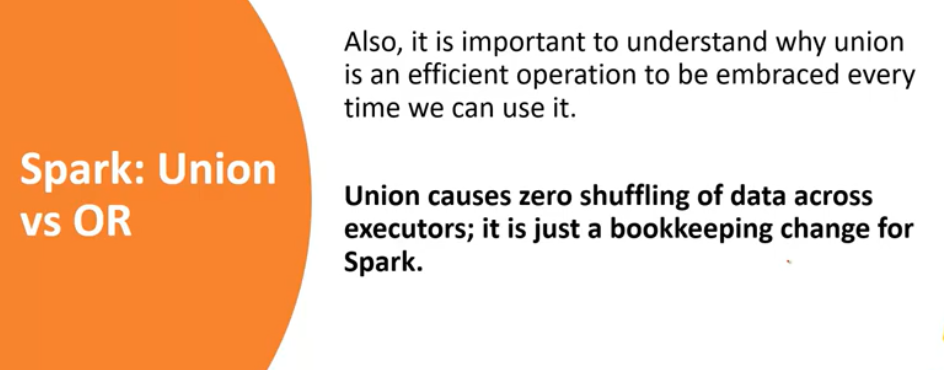


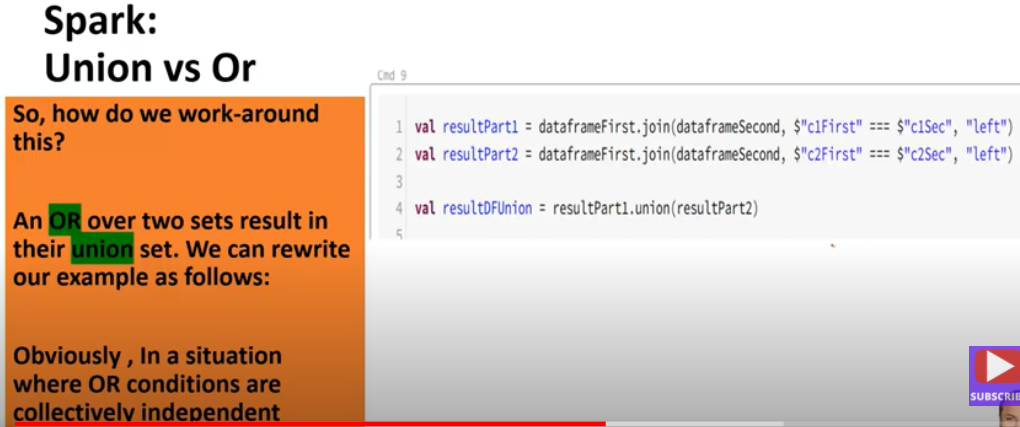
N =no. of record in 1 table

M= no. of records in 2 table



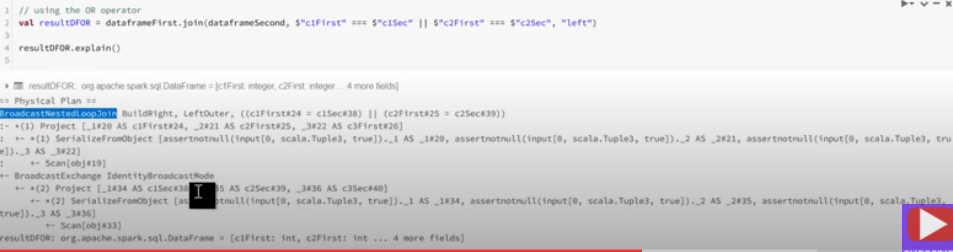






Using or operator in join, result to the broadcast nestedloopjoin, so don’t used that





Instead of this, use union operation which increase your performance and reduce shuffling also. And also result into sortmerge join which is very efficient.

