Column Pruning

Objective

Understand a join optimization done OOTB

Exploit pruning pushdown

Optimize joins with pruning and map-side operation pushdown

Column Pruning

Spark selects just the relevant columns after a join

If you do a select after a join, the Project operation is pushed to joined DFs

Further map-side operations can be manually pushed down

- if we anticipate the joined DF is bigger than either side

Spark sometimes can't prune columns automatically

- good practice: select just the right columns ourselves before join

Most benefits seen in massive datasets

Spark rocks