Big Data Technologies

Trainer: Mr. Nilesh Ghule.

Dataframe Operations

Dataframe Ops

Transformations

$df1 \xrightarrow{op} df2$

① groupBy().agg()
② select()
③ selectExpr()
④ withColumn()
⑤ withColumnRenamed()
⑥ fillna()
⑦ drop()
⑧ limit()
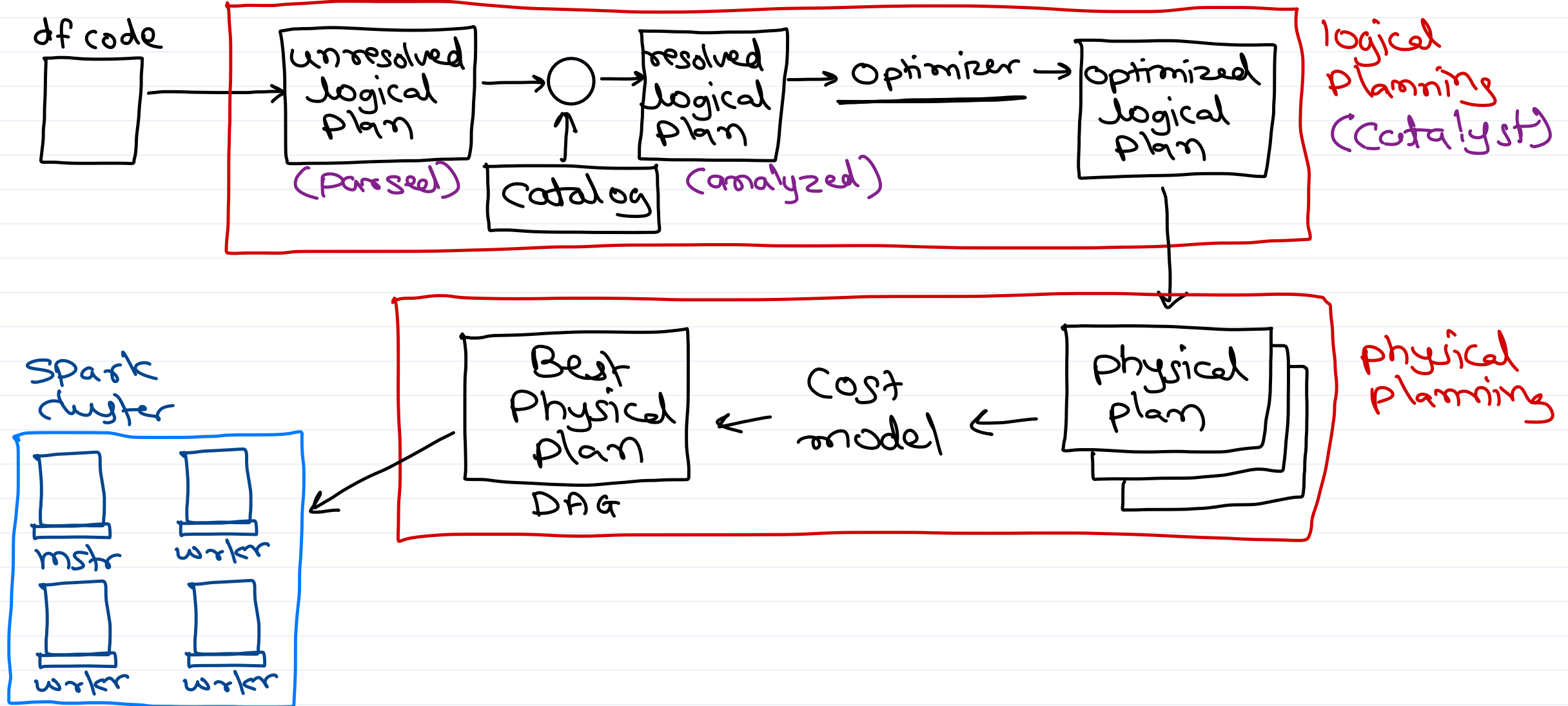⑨ orderBy()
⑩ where() / filter()
⑪ join(), ...

Actions

$df \xrightarrow{op} res$

① show()
② first() → Row
③ collect() → List<Row>
④ take() → List<Row>

---

① read · load()
② write · save()

Dataframe Execution

df code

unresolved logical Plan *(Parsed)*

Catalog

resolved logical Plan *(analyzed)*

Optimizer

Optimized logical Plan

logical Planning (Catalyst)

Best Physical Plan
DAG

Cost model

Physical Plan

physical Planning

Spark cluster

mstr   workr

workr   workr

Thank you!

Nilesh Ghule <nilesh@sunbeaminfo.com>