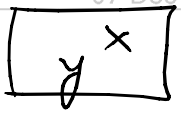
DSM

Note Title 07-Dec-22

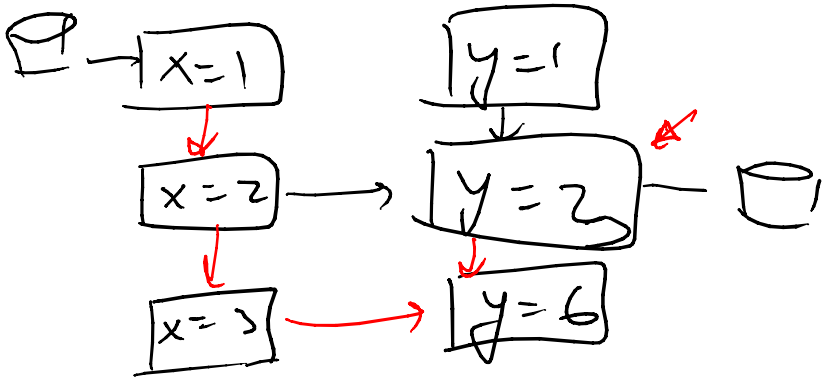
- False sharing



for
$x \leftarrow x+1$
$y \leftarrow y+1$

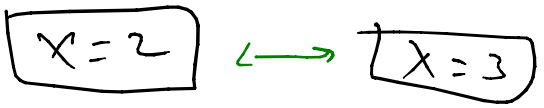
- FT: Checkpoints: Inconsistency

for
$x++$
$y \leftarrow y+x$

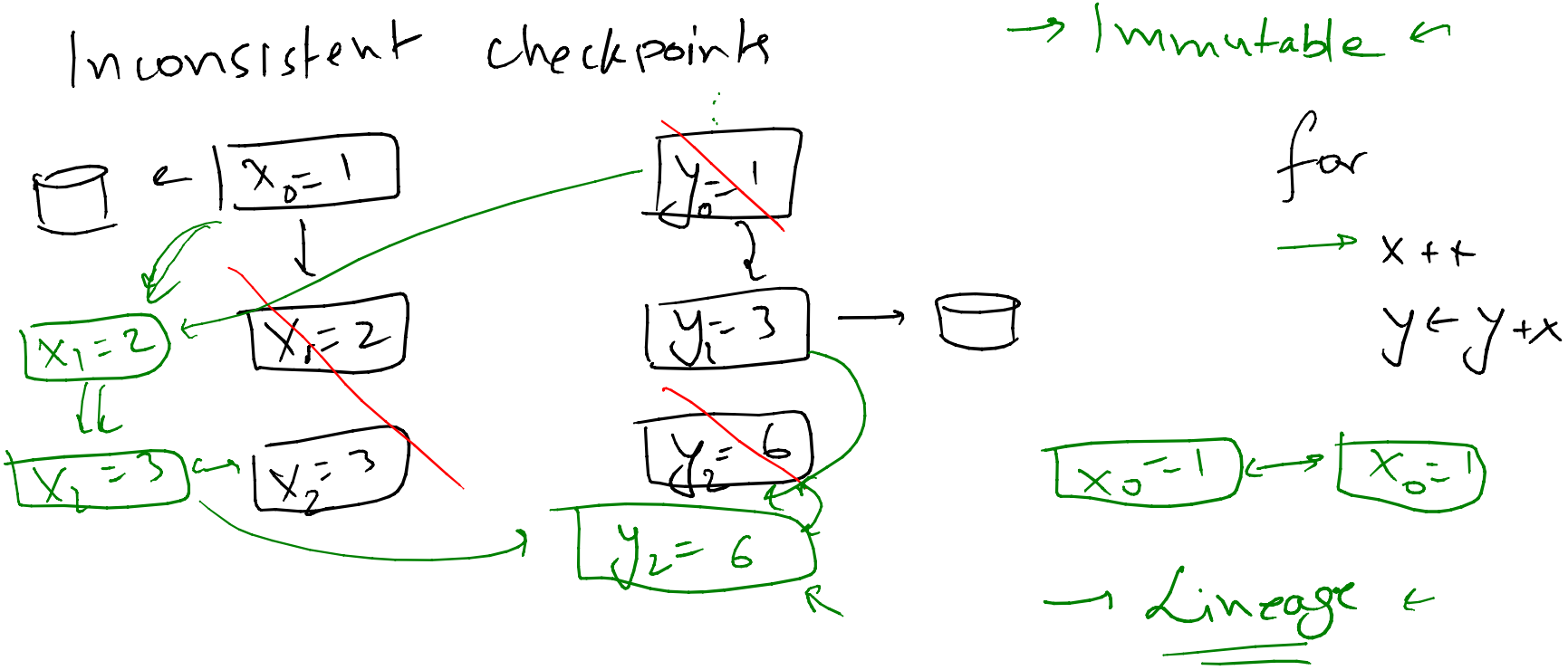


x=2 y=5

Replication - Slow



MapReduce Locality / FT → deterministic ← Idempotence


Inconsistent checkpoints
→ Immutable ←
for
→ x++
y ← y+x
$x_0 = 1$
$y_0 = 1$
$x_1 = 2$
$x_1 = 2$
$y_1 = 3$
$x_2 = 3$
$x_2 = 3$
$y_2 = 6$
$y_2 = 6$
$x_0 = 1$
$x_0 = 1$
→ Lineage ←

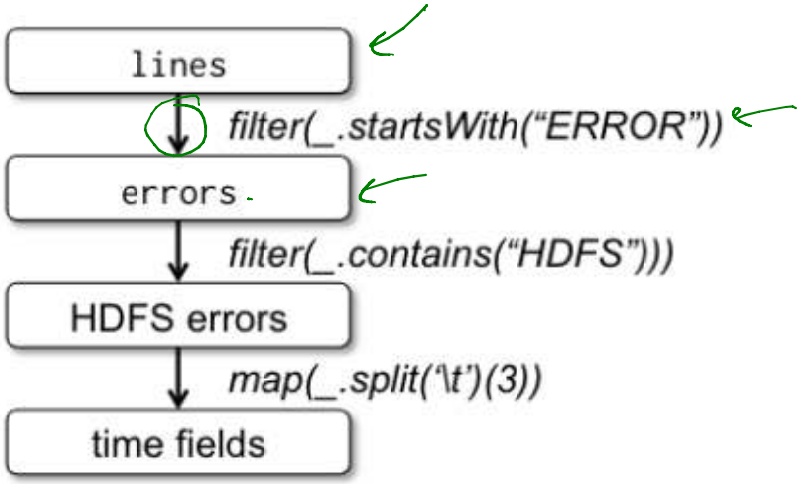
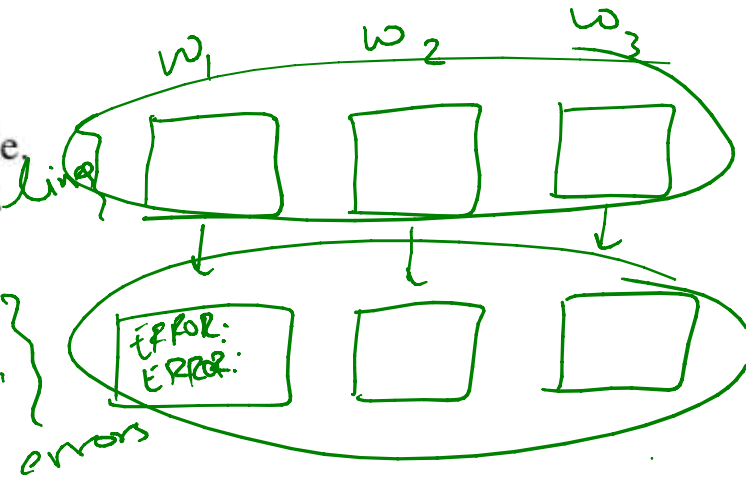
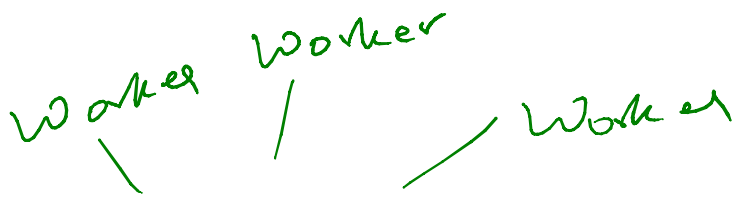


Figure 1: Lineage graph for the third query in our example. Boxes represent RDDs and arrows represent transformations.

```
lines = spark.textFile("hdfs://...")
errors = lines.filter(_.startsWith("ERROR"))
errors.persist()
```

Resilient Distributed Datasets

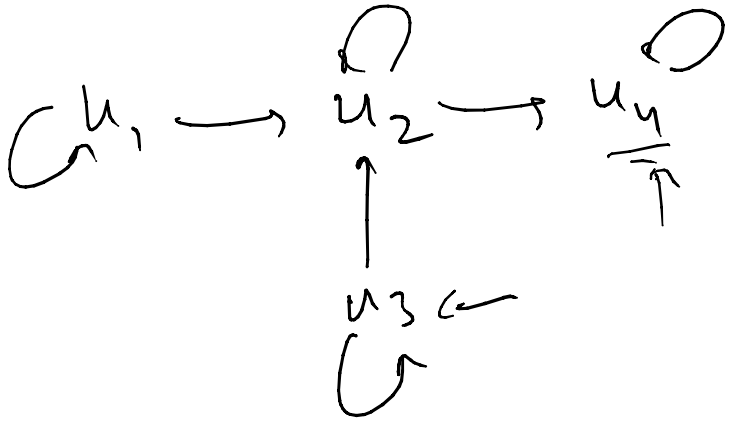




RDDs — partitions

P — worker

Page Rank



$u_1, u_3 < u_2$

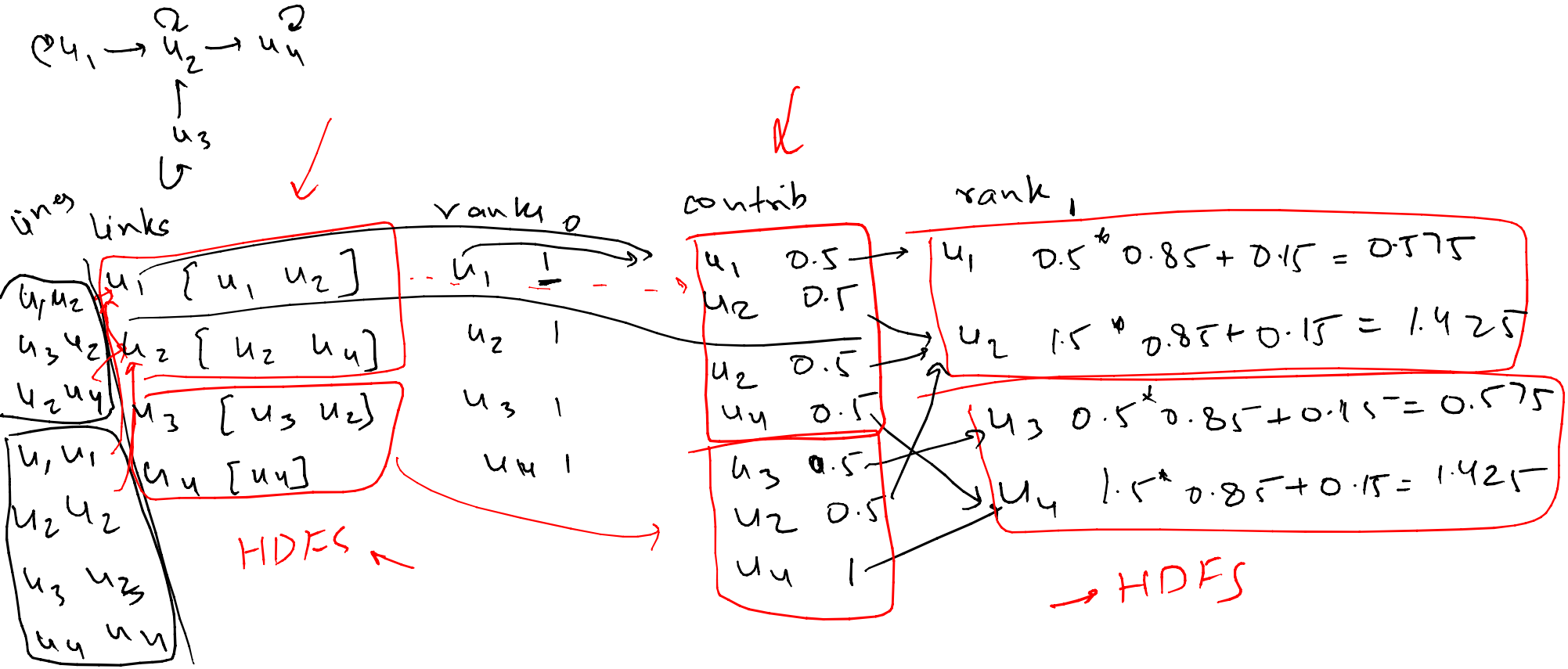
$u_1, u_3 < u_4$

$u_2 < u_4$

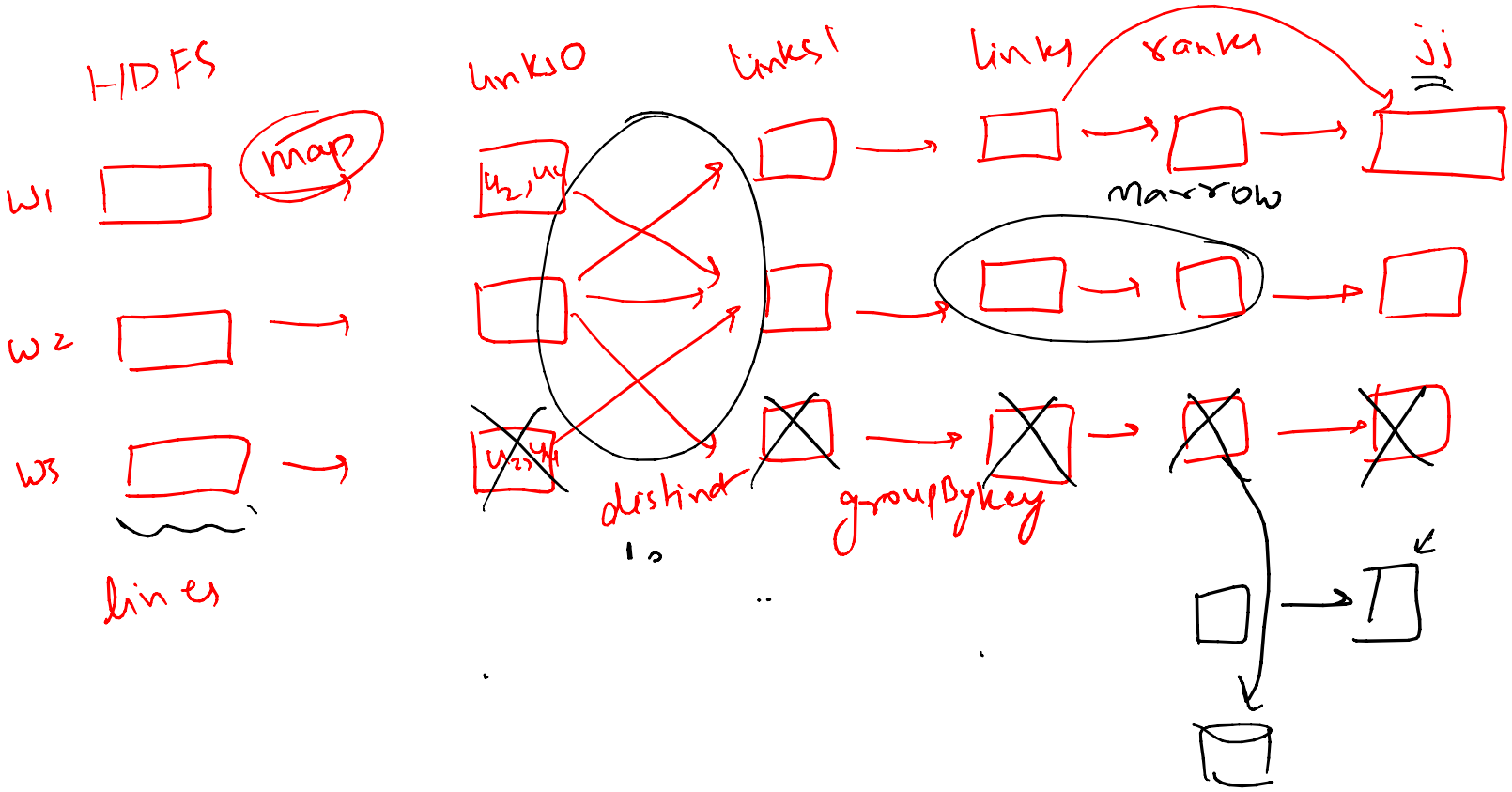
0.15

0.85

u_1



u1 → u2 → u4
u3
lines
links
rank 0
contrib
rank 1
u1 u2
u3 u2
u2 u4
u1 u1
u2 u2
u3 u3
u4 u4
u1 [u1 u2]
u2 [u2 u4]
u3 [u3 u2]
u4 [u4]
u1 1
u2 1
u3 1
u4 1
u1 0.5
u2 0.5
u2 0.5
u4 0.5
u3 0.5
u2 0.5
u4 1
u1 0.5*0.85+0.15 = 0.575
u2 1.5*0.85+0.15 = 1.425
u3 0.5*0.85+0.15 = 0.575
u4 1.5*0.85+0.15 = 1.425
HDFS
HDFS


HDFS
w1
w2
w3
lines
map
links0
links1
links
ranks
jj
marrow
distinct
groupBykey

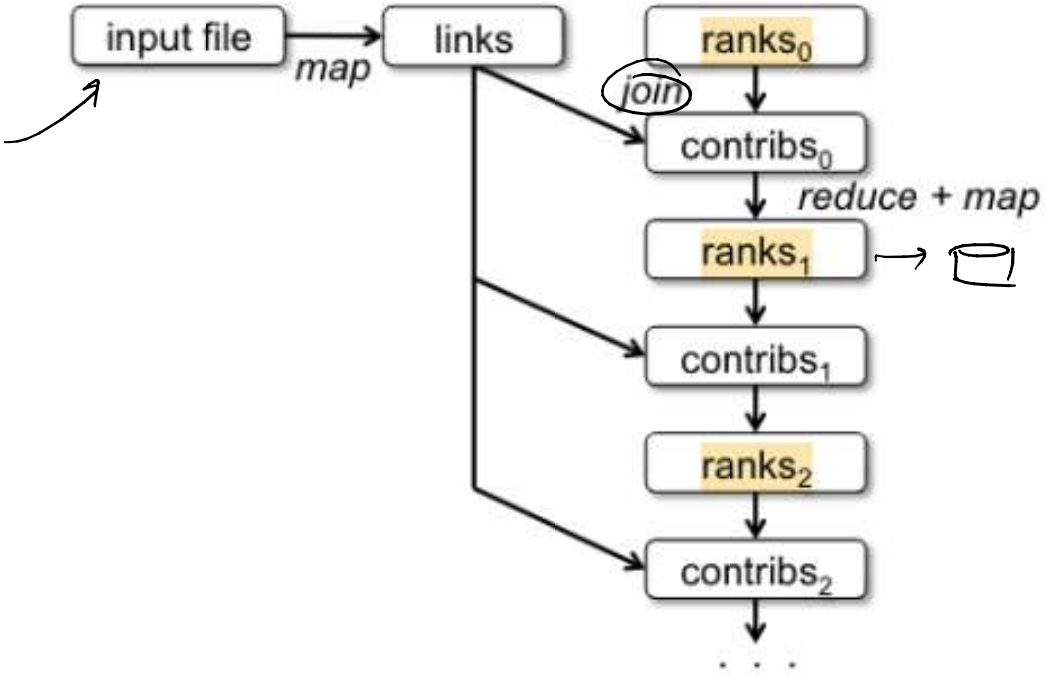
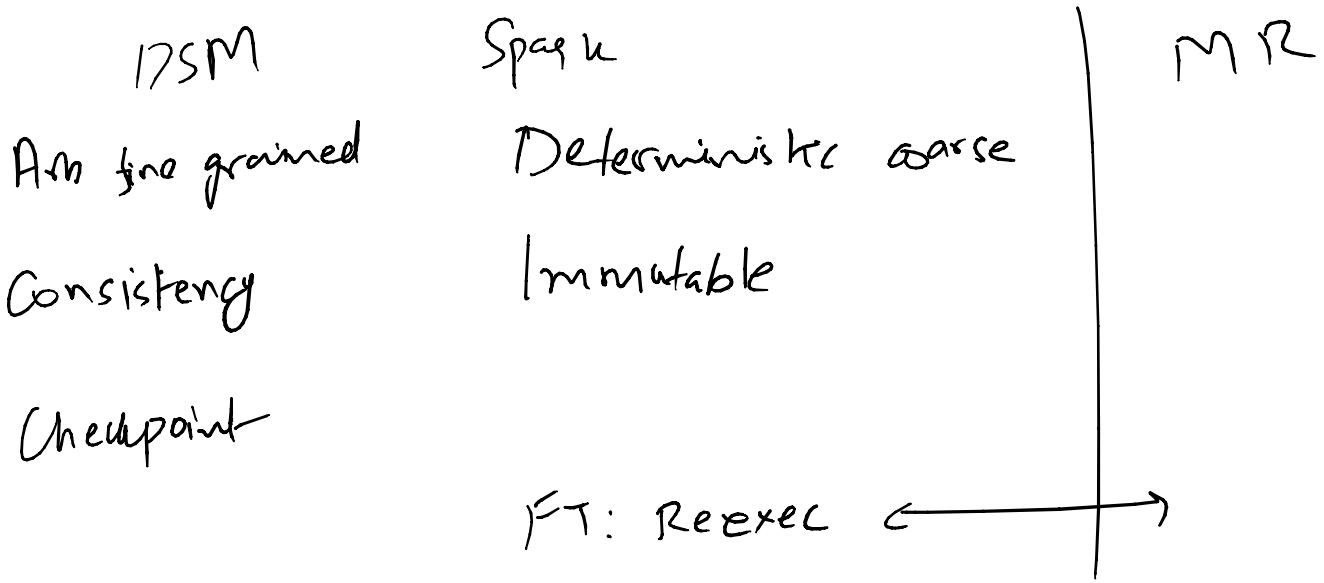


Figure 3: Lineage graph for datasets in PageRank.


ISM
Spark
MR
Arb fine grained
Deterministic coarse
Consistency
Immutable
Checkpoint
FT: Reexec