RDD – (Resilient Distributed Dataset).

Resilient → Fault tolerance.

Distributed → spread across cluster

Dataset → collections of data.

Immutable →

recompute → RDD – lineage graph (DAG). → RDD1 – text file

RDD2 = filter rdd

~~RDD3~~ – map function

Apache Spark is an in-memory data processing framework.

in-memory → RAM.

① Spark content - (Entry point to the spark functionality).

Narrow - No shuffle operations.

wide - there will be shuffle operations.

Catalyst optimister and tungsten execution engine

RDD → DF → ~~DS~~ (Not in pyspark)
(scala)

→ Do read about it

~~RDD → Dataset~~

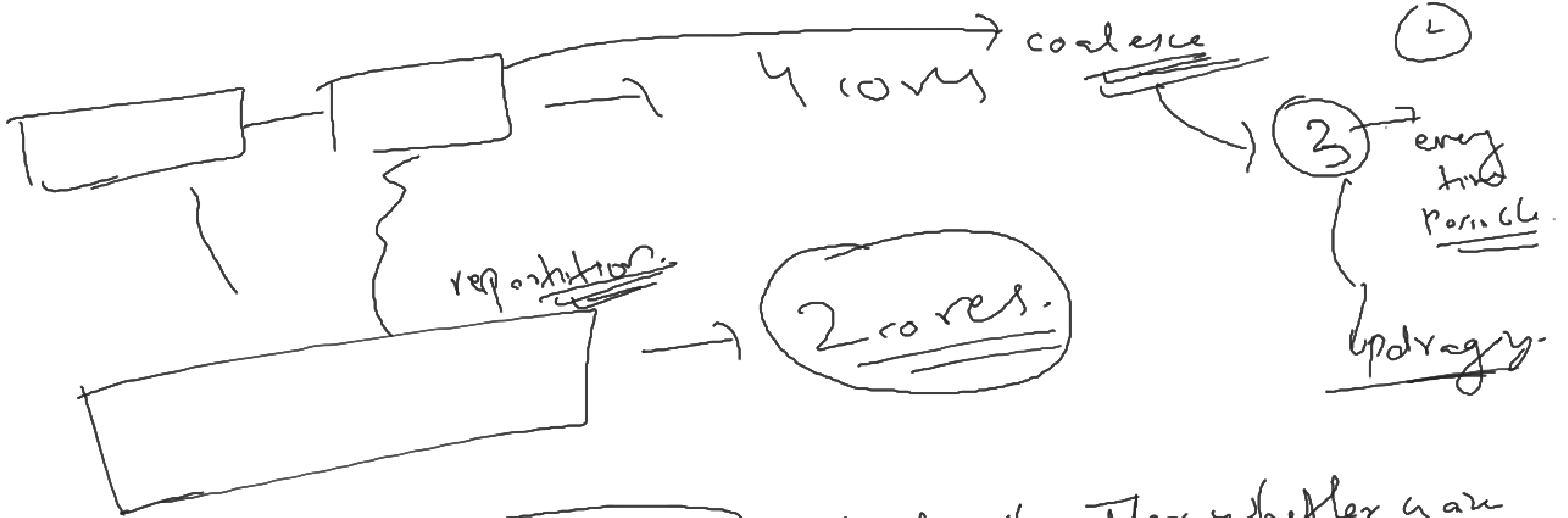
Hyper Threading in OS.

1) Quad core - (4 cores) → Programm running to behave by hardware.

octacore (   )

smartphones - ① octacore - (Hyper Threading)

② octacore - by hyperthreading



Coalesce – minimum shuffling (It doesn't matter whether u are

increasing | decreasing

repartition → maximum shuffling → consequences into your consideration

2 → 8