**Problem statement:**

Visualise how data is distributed as RDD across a cluster of 5 machines of 4 cores. Describe a complete picture of RDD, partitions, lambda implementing on each partition and correlation with the JVMs.

**Walkthrough:**

RDD’s are:

Resilient - Fault tolerant (immutable)

Distributed – Divided across the clusters

Datasets – collection of records.

The Spark automatically partitions RDDs and distribute the partitions across different nodes. The RDDs and its partitions define the basic units of parallelism in spark.

In local mode the default no. of partitions is the number of cores allotted to the application.

While in cluster mode, the no. of partitions depends on no. of blocks or input splits.

So by default parallelism in spark is no. of cores allocated in sc.

**Snapshots:**

I have extracted the image behind executors and partitions in relation with JVM, the options to check micropartitions and internal functioning through databricks is not proper , do have a look at snapshots and guide me.

