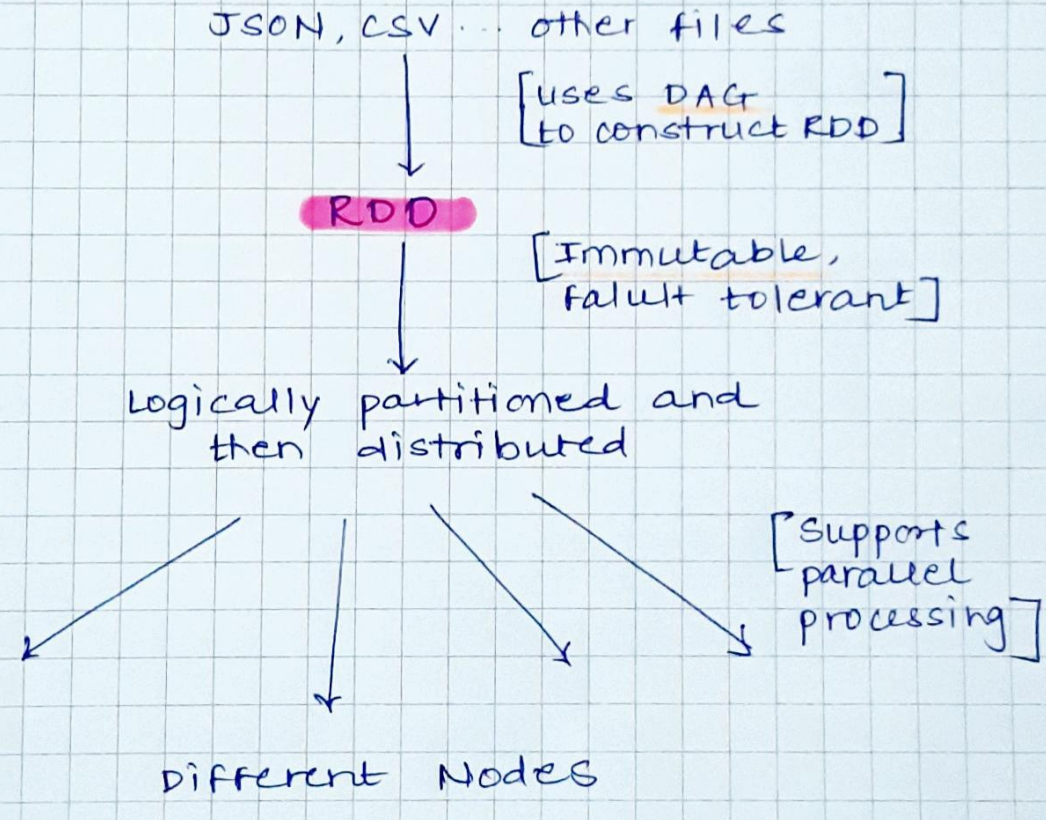
SPARK RDD

RDD : Resilient Distributed Dataset

- Fault Tolerant
- We can recompute the lost/missing/corrupted data/partition using RDD.

- Every dataset is logically partitioned and distributed across servers.
- The data structure used for logical distribution is RDD.



Note : DAG - Directed Acyclic Graph.

Spark RDD Features

	Features	Brief about feature.
1.	In-memory computation	• RDDs are stored in-memory for desired timeline
2.	Lazy Evaluation	• computation is performed only after action is triggred
3.	Fault Tolerent	• In case of loss/failure the partition can be recomputed.
4.	Immutable	• Once created, RDDs cant be manipulated.
5.	Persistance	• Frequently used RDD's can be stored in in-memory.
6.	Partitioning	• seamless distribution on nodes in the cluster.
7.	Parallelism	• RDDs are processed parallely.
8.	Typed	• like RDD[int, long, string]
9.	No limitation	• Supports any number of RDDs.
10.	Coarse-grained	• Coarse-grained: operations gets applied to entire RDD.
11.	Location-stickiness	• Data is placed closer to task using placement reference.

RDD Operations
1. Transformation
2. Action.

1. Transformation :

parent-RDD
↓
Transformation
↓
New-RDD.

- RDD (Resilient Distributed Dataset)
- Transformation is lazy operation.

Transformation
- Wide
- Narrow

Narrow-Transformation

Parent RDD

Partition 1 . . . . Partition N.

↓ ↓

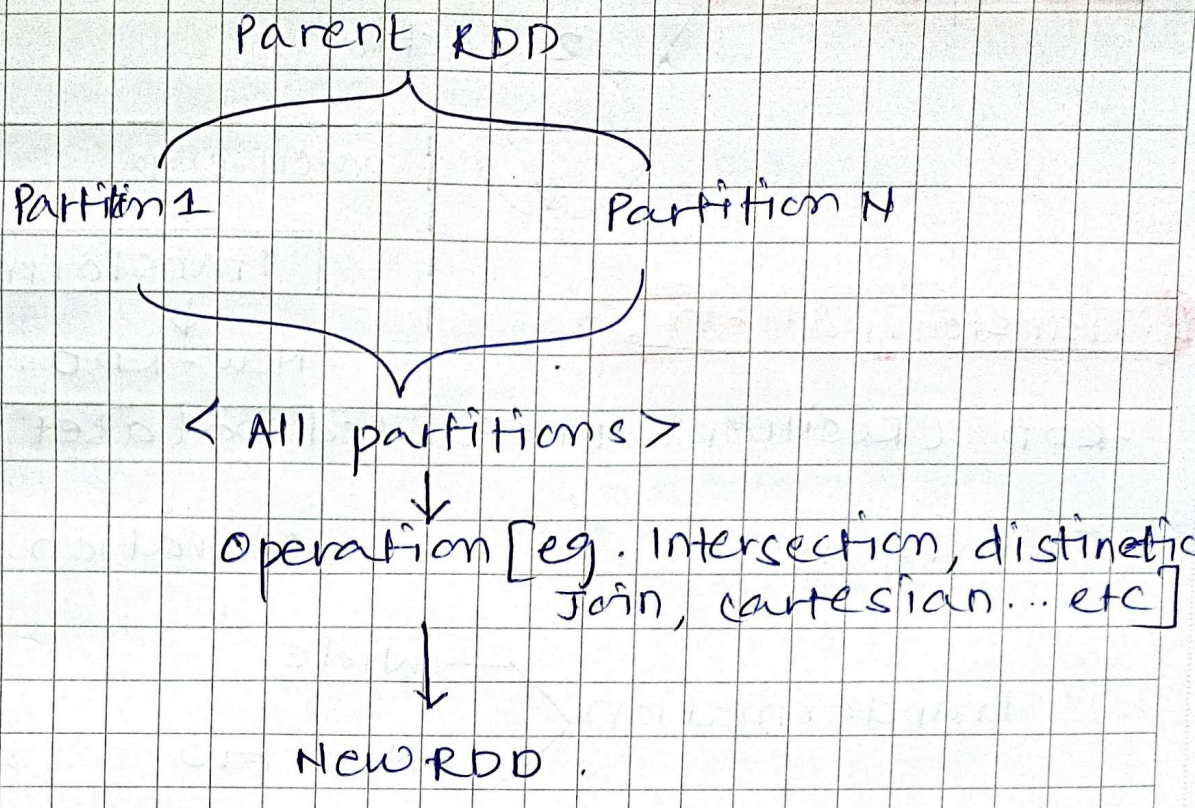
< some partitions >

↓

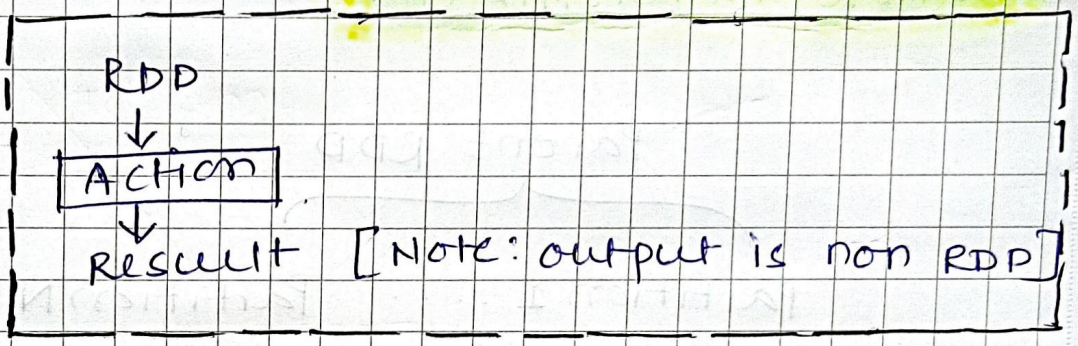
Operation [eg. FlatMap, Map, Filter, Sample

↓

New-RDD.



2. Action :



- Few operations under 'Action' can be First(), take(), count() etc.

Spark: Map and FlatMap:

- What is Map?

Takes '1' entity as an input and generates exactly 1 entity as a output.

- What is FlatMap?

Takes '1' entity as a input and generates '0' or 'many' entities as a output.

In Spark, Map and FlatMap are transformations, hence input and output generated are RDDs.

Map

RDD

↓

Map programming_logic()

↓

RDD

#Size of input RDD and output RDD are same.

Flat-Map

RDD

↓

FlatMap Programming_logic()

↓

0 or N-RDD

Size of input and output RDD is different.