What are Pair RDDs

Pair RDDs are those RDDs that contain **key and value** pairs

Key1, Value1

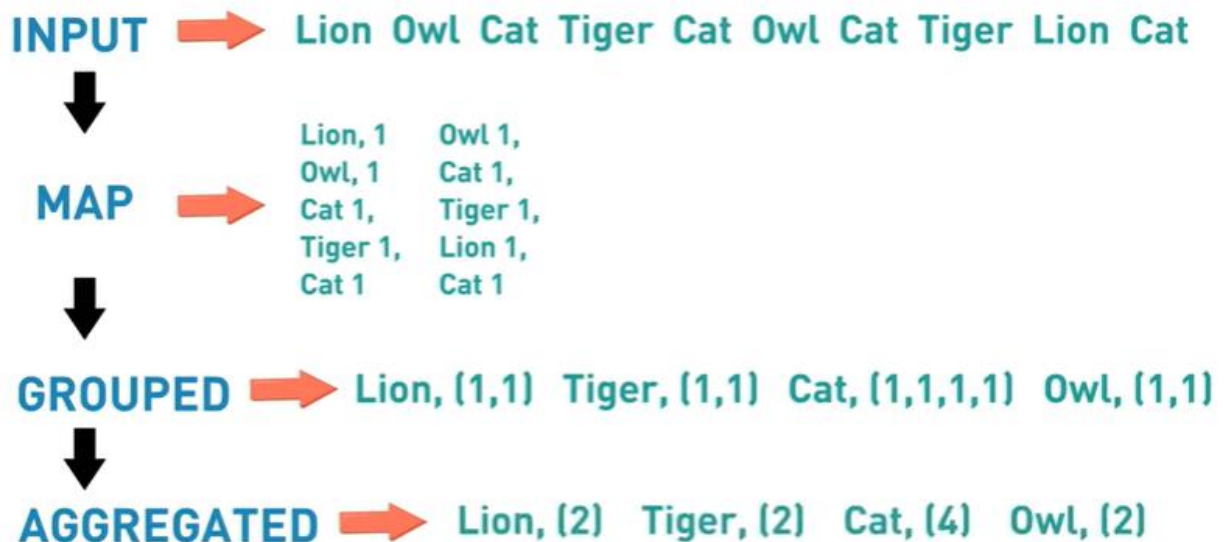**Key2, Value2**

.

.

KeyN, ValueN

Why do we need Pair RDDs?

Why do we need Pair RDDs

INPUT →	Lion Owl Cat Tiger Cat Owl Cat Tiger Lion Cat

↓

MAP →	Lion, 1 Owl 1, Owl, 1 Cat 1, Cat 1, Tiger 1, Tiger 1, Lion 1, Cat 1 Cat 1

↓

GROUPED → Lion, (1,1) Tiger, (1,1) Cat, (1,1,1,1) Owl, (1,1)

↓

AGGREGATED → Lion, (2)   Tiger, (2)   Cat, (4)   Owl, (2)

Why do we need Pair RDDs

These key-value RDDs are instrumental in use cases where we need to group and/or aggregate values based on a key. When we create pair RDDs, spark allows us to apply certain special key-value operations like

- groupByKey
- reduceByKey