Introduction to Map Reduce

@void_io

Introduction to Map Reduce

What is MapReduce

MR Phases

MR Key-Value

MR SDLC

MR Process

MR Design

What is Map Reduce?

Hadoop MapReduce is a software framework for writing applications which process vast amounts of data in-parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner.

Map Reduce Phases

MapReduce programs work in two phases:

- Map phase
- Reduce phase

The MapReduce framework operates exclusively on <key, value> pairs, that is, the framework views the input to the job as a set of <key, value> pairs and produces a set of <key, value> pairs as the output of the job, conceivably of different types.

Map Reduce <Key, Value>

- **TextInputFormat(default)**
  - Key: offset(system generated)
  - Value: rest of the data
- **SequenceInputFormat**
  - Key: sequence number
  - Value: rest of the data
- **FileInputFormat**
  - Key: upto 1st delimeter
  - Value: rest of the data

Map Reduce SDLC

- Requirement Gathering
- MR Design
- Developemt
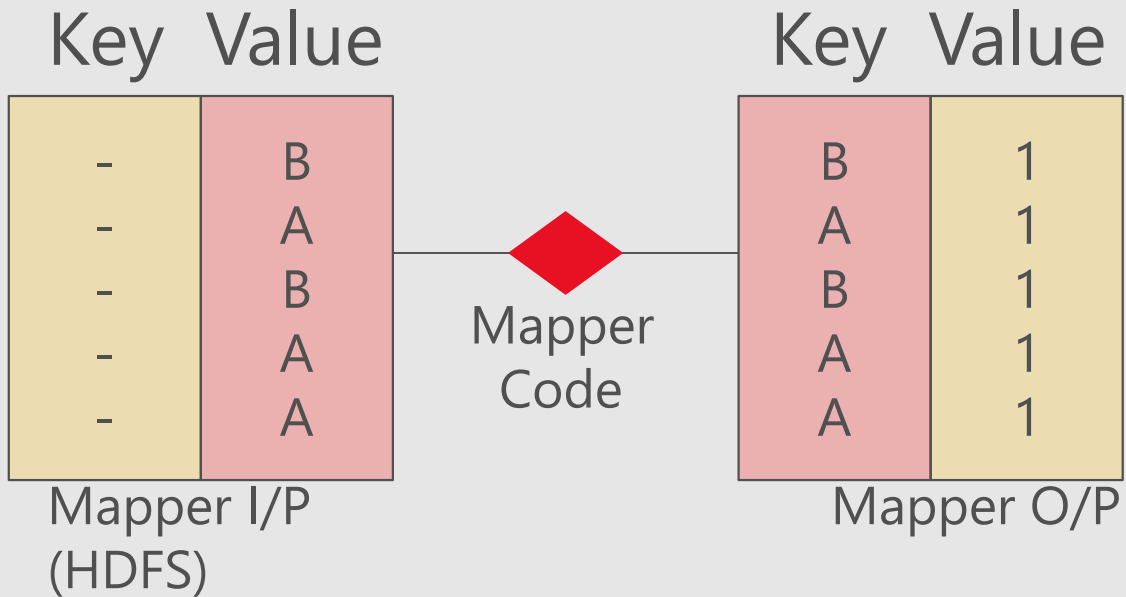- Testing
- Deployment(production environment)
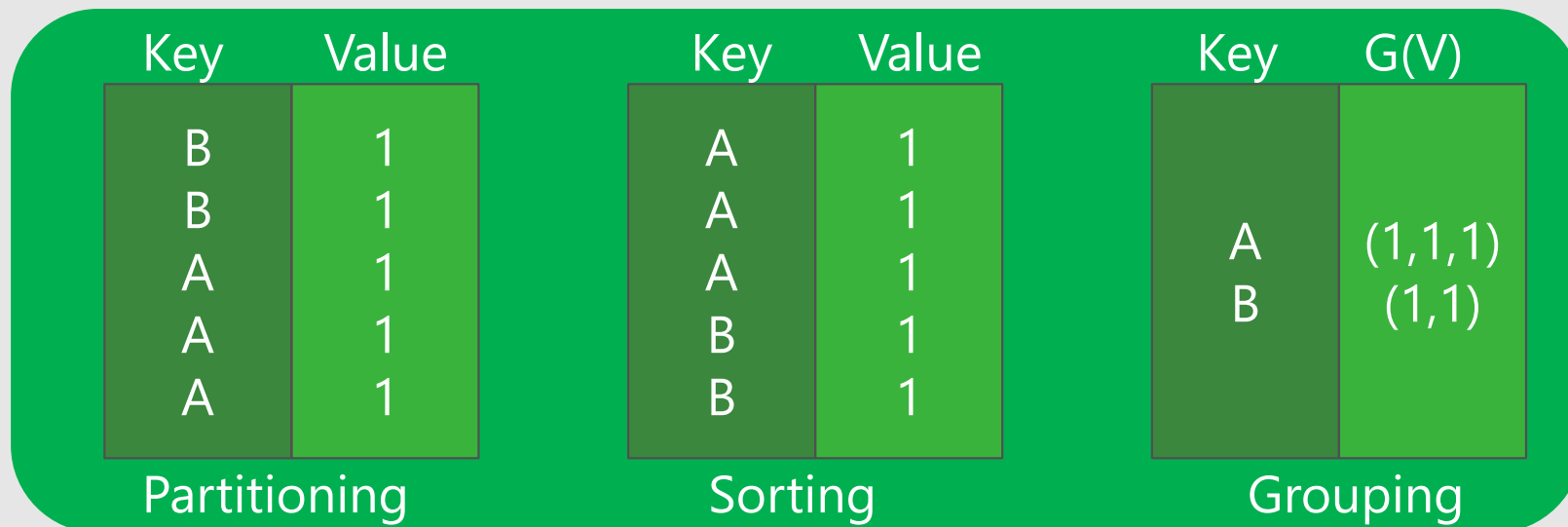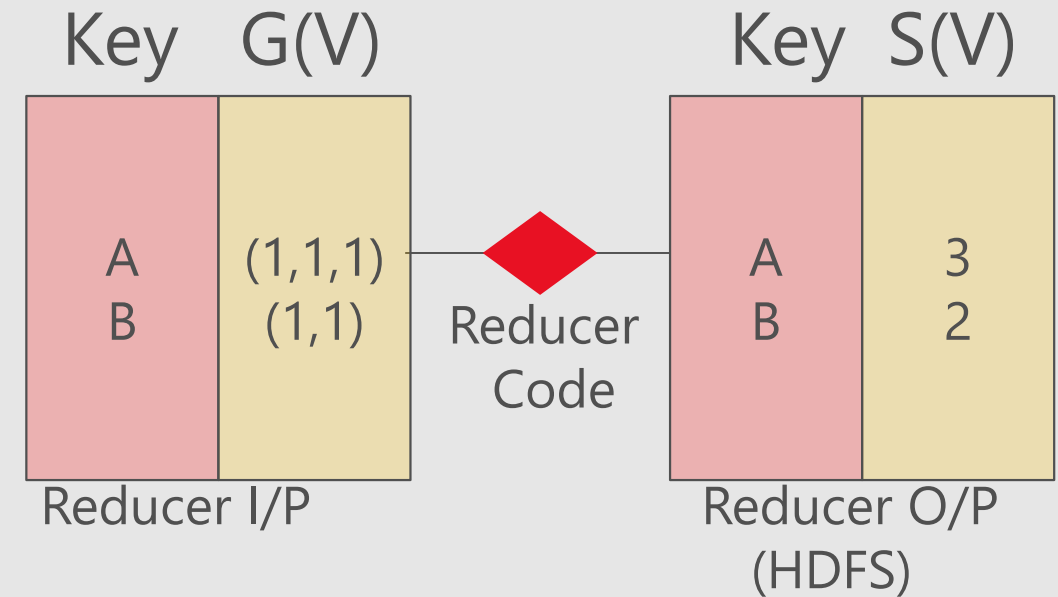
MR Process

- Mapper Task

  Combiner Task
  Partitioning Task(Hash)
  Sorting Task (Ascending)
  Grouping Task
  Shuffle Task

- Reducer Task

|__Mapper Task____| |__Reducer Task___|

Key	Value
- - - -	B A B A A

Mapper I/P (HDFS)

Mapper Code ◆

Key	Value
B A B A A	1 1 1 1 1

Mapper O/P

COMBINER | **FRAMEWORK**

Key	G(V)
A B	(1,1,1) (1,1)

Reducer I/P

Reducer Code ◆

Key	S(V)
A B	3 2

Reducer O/P (HDFS)

Key	Value
B B A A A	1 1 1 1 1

Partitioning

Key	Value
A A A B B	1 1 1 1 1

Sorting

Key	G(V)
A B	(1,1,1) (1,1)

Grouping

Thank you ☺

@void_io