

# Map-Reduce

MIE1628 office hours

**Objective:** find out total fees collected by each department

Student ID	Department	Fee
1	Mech	10000
6	CS	20000
7	Elec	5000
2	Mech	12000
4	Mech	16000
3	Elec	18000
9	CS	16000
5	Mech	6000

Input

Output

Department	Fee
CS	36000
Elec	23000
Mech	44000

# Map-Magic-Reduce

## Map

```
1, Mech, 10000
6, CS, 20000
7, Elec, 5000
2, Mech, 12000
4, Mech, 16000
3, Elec, 18000
9, CS, 16000
5, Mech, 6000
```

Node 1

```
1, Mech, 10000
6, CS, 20000
7, Elec, 5000
```

```
Mech, 10000
CS, 20000
Elec, 5000
```

Node 2

```
2, Mech, 12000
4, Mech, 16000
3, Elec, 18000
```

```
Mech, 12000
Mech, 16000
Elec, 18000
```

Node 3

```
9, CS, 16000
5, Mech, 6000
```

```
CS, 16000
Mech, 6000
```

Magic

## Reduce

Node 1

```
CS, [20000, 16000]
Elec, [5000, 18000]
Mech, [10000, 12000, 16000, 6000]
```

```
CS, 36000
Elec, 23000
Mech, 44000
```

# Map

## Map

Node 1

1, Mech, 10000
6, CS, 20000
7, Elec, 5000

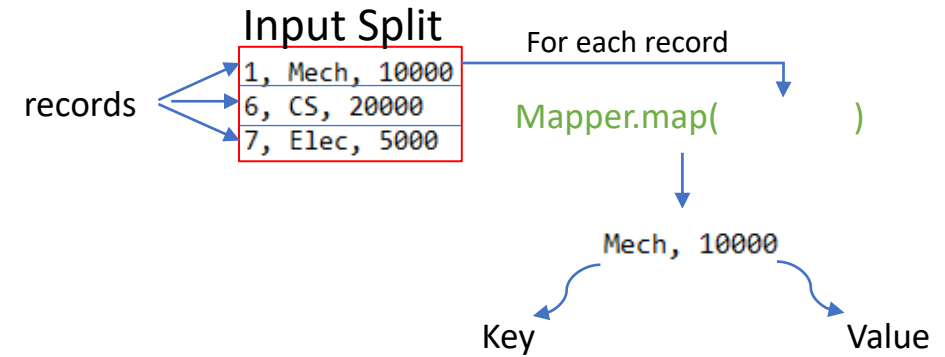
Node 2

2, Mech, 12000
4, Mech, 16000
3, Elec, 18000

Node 3

9, CS, 16000
5, Mech, 6000

1, Mech, 10000
6, CS, 20000
7, Elec, 5000
2, Mech, 12000
4, Mech, 16000
3, Elec, 18000
9, CS, 16000
5, Mech, 6000

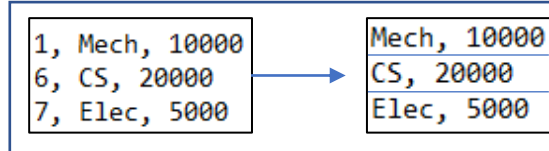


For each input split, a mapper object is created by the Hadoop MR framework using the Mapper class you provide. The map method of that object is called for each record in the split. The map method should output a key-value pair.

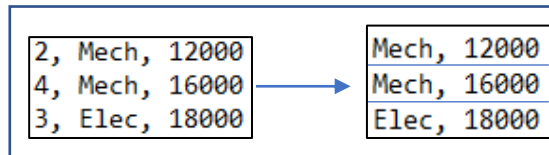
# Map

## Map

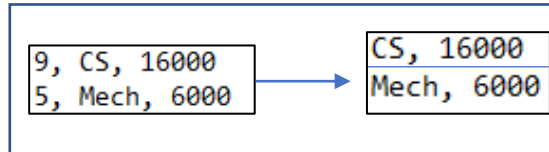
Node 1



Node 2



Node 3



1, Mech, 10000
6, CS, 20000
7, Elec, 5000
2, Mech, 12000
4, Mech, 16000
3, Elec, 18000
9, CS, 16000
5, Mech, 6000

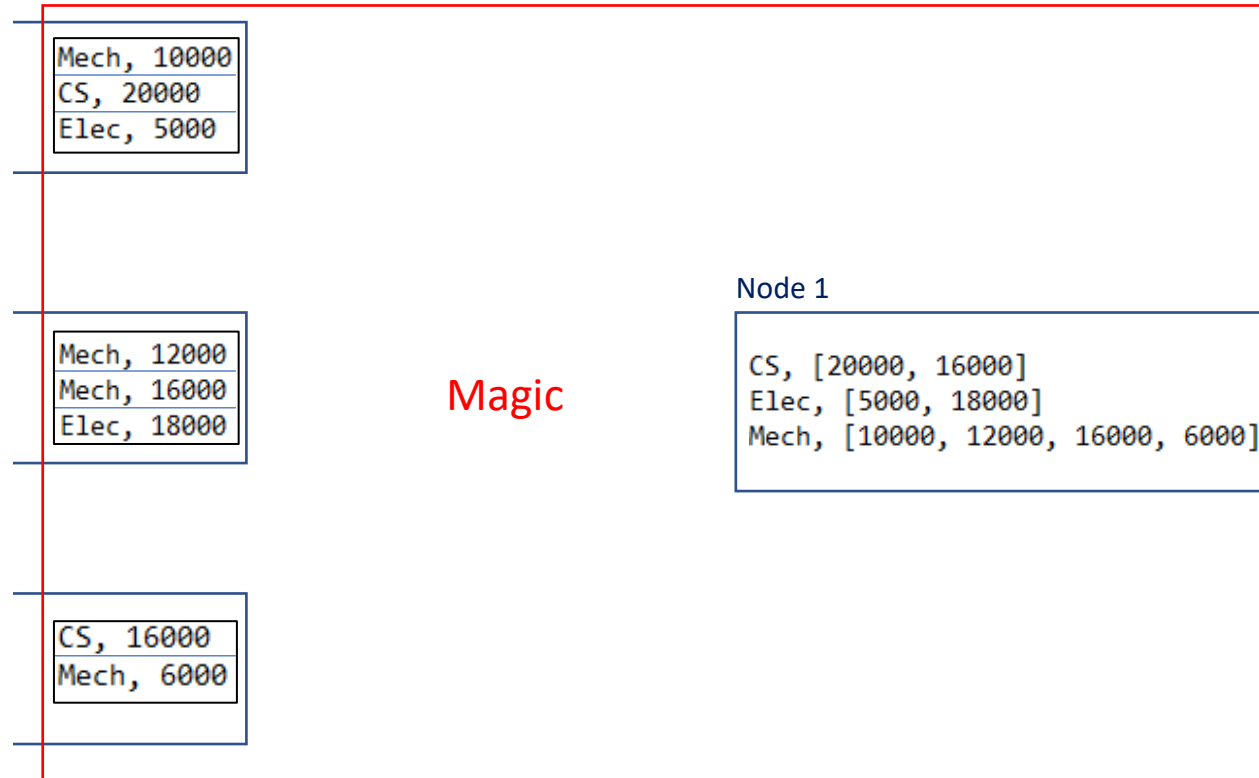
# Magic

Between Map and Reduce, the Hadoop MR framework shuffles the output of Map (k-v pairs) across nodes. The pairs with the same key are transmitted to the same node. They are sorted by key and the values corresponding to each key are grouped into a list.

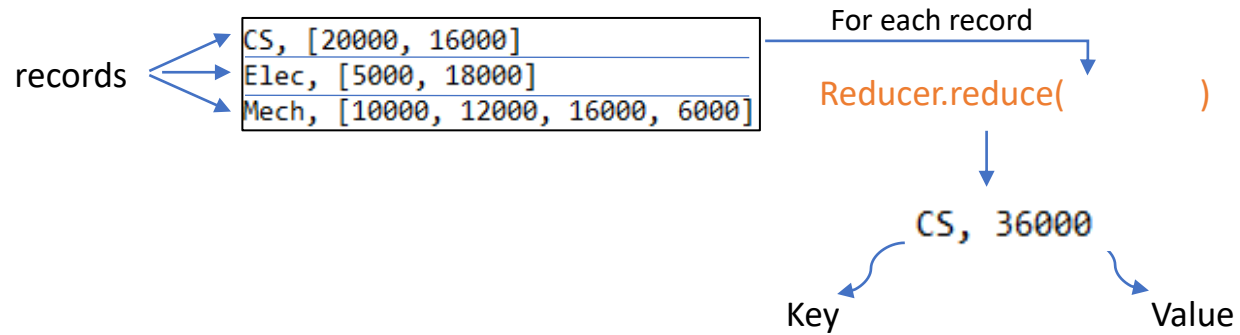
The sorting and grouping is carried out on both Mapper and Reducer sides for efficiency, and the actual process is much complex and opaque. The number of nodes used on the reducer side depends on the number of reducer used.

The programmers do not need to do anything. The framework takes care of everything. **Only note that the reducer's input will be a pair of a key and a list of corresponding values.**

Sometime, the combiner is used on the mapper side before shuffling. It acts like mini reducer, that is, it reduces a list of values into a value for each key. It helps in reducing the amount of data to be transmitted over the network. It also reduces the workload of reducer by performing the same operation.



# Reduce



A reducer object is created by the Hadoop MR framework using the Reducer class you provide. The reduce method of that object is called for each record assigned to that reducer node. The reduce method should output a key-value pair.

