OutputFormat

Validate output specifications

RecordWriter implementation to write output files of the job

InputFormat

Validate inputs

Input files into logical InputSplits

RecordReader implementation to extract logical records

REDUCER

- Reduce function will take Key Value pairs from multiple Map functions as input and Reduce them in output

- Keys are grouped with values. Reduce function is called once per key and its values.

- There could be 0, 1 or more Reduce function for a MapReduce job
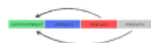
DISSECTING MAPREDUCE PROGRAM

Writable

A serializable object which implements a simple, efficient, serialization protocol

Fast, compact & effective

MAPPER

- Dataset is divided in to multiple parts - Input Splits

- Each Mapper process an Input Split

- Each Mapper can be called multiple times depending on the content of Input Split

- Mapper will emit Key Value pairs as output
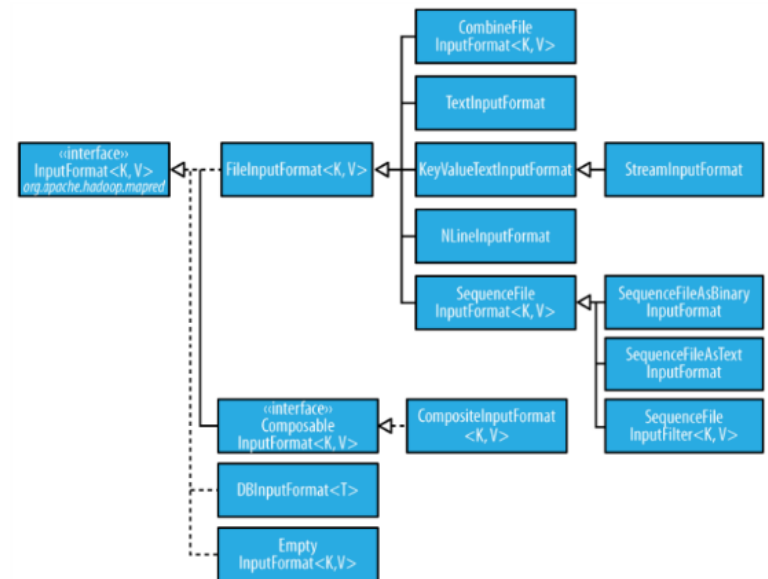
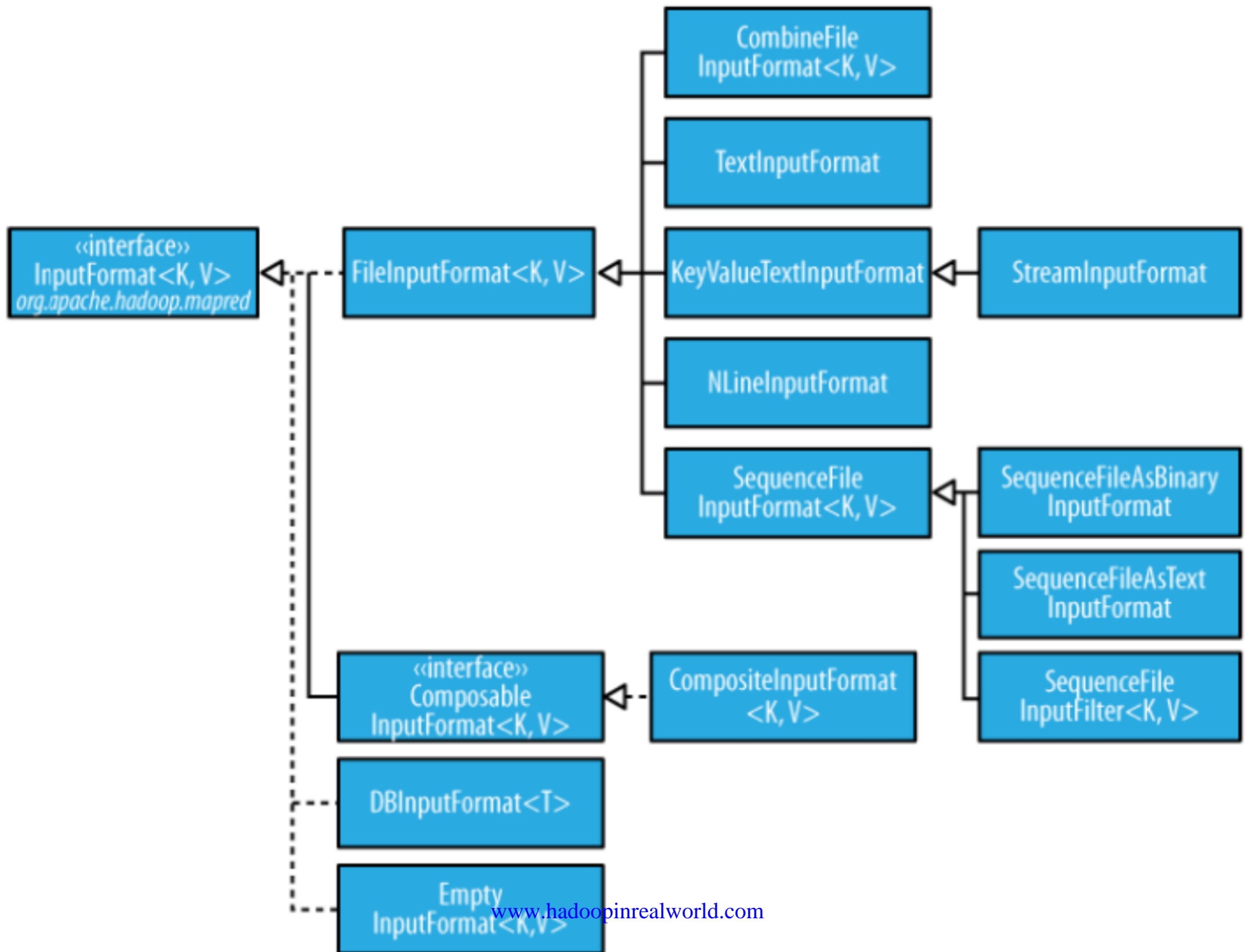- There will be one or more Mapper in a MapReduce job

DISSECTING MAPREDUCE PROGRAM

InputFormat

Validate inputs

Input files into logical InputSplits

RecordReader implementation to extract logical records

```
                                                    ┌──────────────────────┐
                                                    │    CombineFile        │
                                                    │  InputFormat<K, V>    │
                                                    └──────────────────────┘

                                                    ┌──────────────────────┐
                                                    │   TextInputFormat     │
                                                    └──────────────────────┘

┌──────────────────────┐      ┌──────────────────────┐   ┌──────────────────────┐   ┌──────────────────────┐
│     «interface»       │      │  FileInputFormat<K, V>│   │ KeyValueTextInputFormat│  │   StreamInputFormat   │
│  InputFormat<K, V>    │◁─ ─  └──────────────────────┘   └──────────────────────┘   └──────────────────────┘
│ org.apache.hadoop.mapred│
└──────────────────────┘

                                                    ┌──────────────────────┐
                                                    │   NLineInputFormat    │
                                                    └──────────────────────┘
```

SequenceFile InputFormat<K, V>

SequenceFileAsBinary InputFormat

SequenceFileAsText InputFormat

SequenceFile InputFilter<K, V>

«interface» Composable InputFormat<K, V>

CompositeInputFormat <K, V>

DBInputFormat<T>

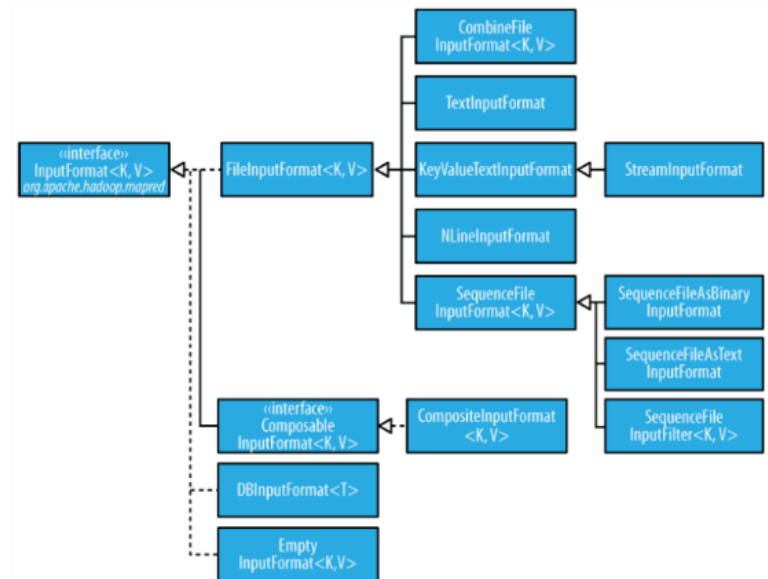Empty InputFormat<K, V>

www.hadoopinrealworld.com

InputFormat

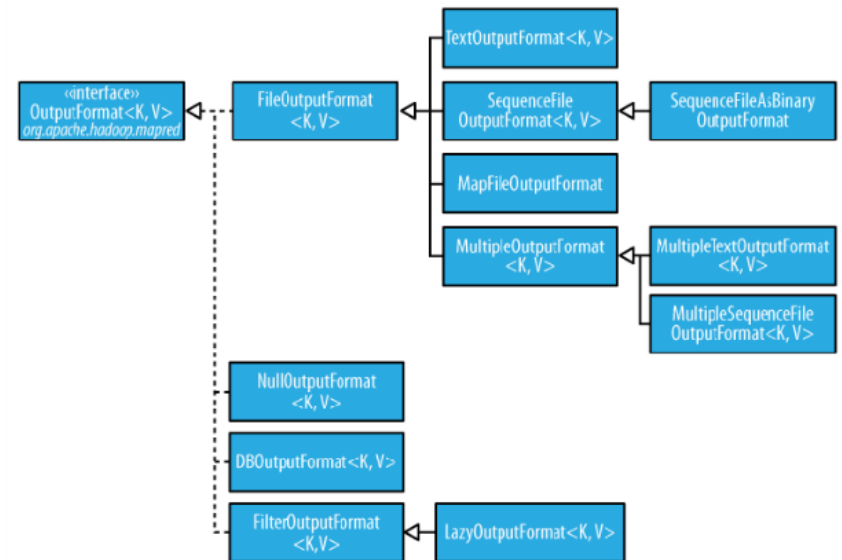Validate inputs

Input files into logical InputSplits

RecordReader implementation to extract logical records

OutputFormat
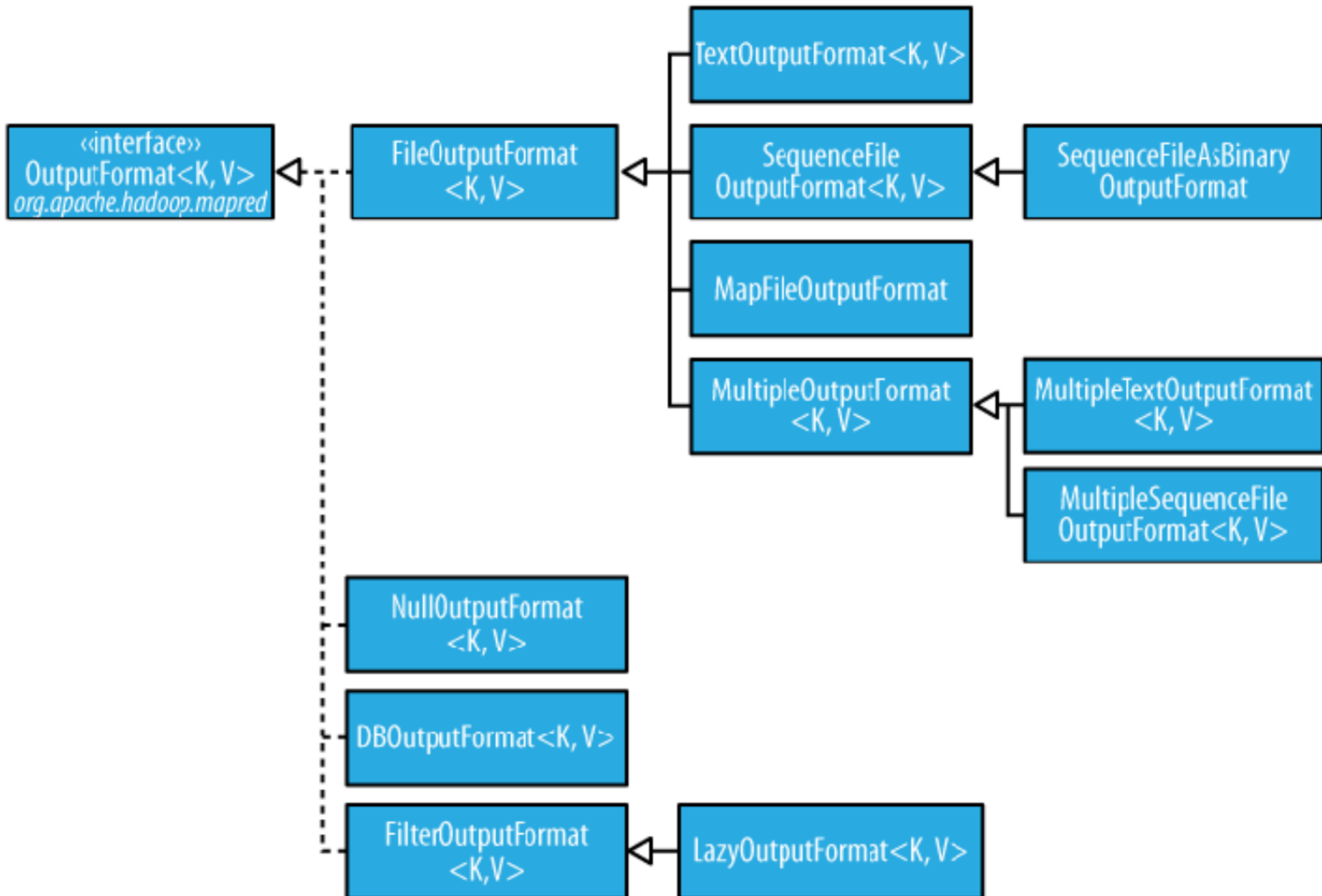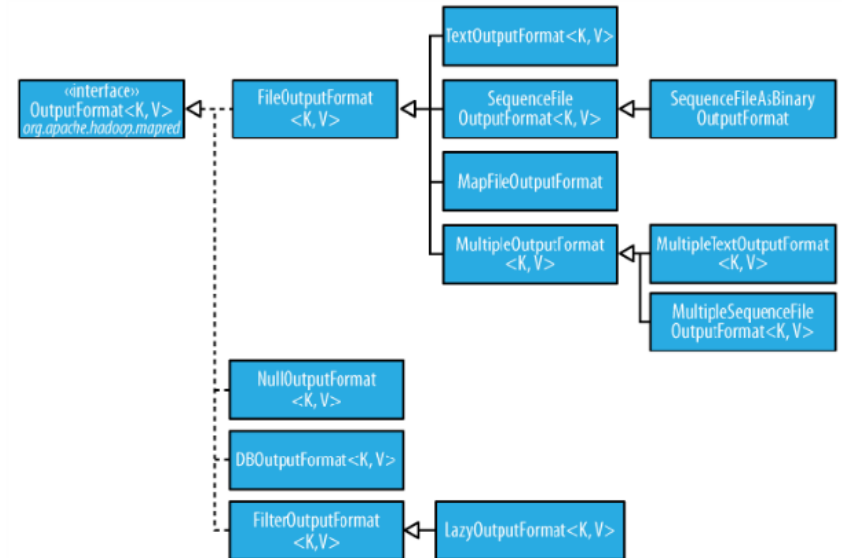
Validate output specifications

RecordWriter implementation to write output files of the job

```
«interface»
OutputFormat<K, V>
org.apache.hadoop.mapred

FileOutputFormat
<K, V>

TextOutputFormat<K, V>

SequenceFile
OutputFormat<K, V>

SequenceFileAsBinary
OutputFormat

MapFileOutputFormat

MultipleOutputFormat
<K, V>

MultipleTextOutputFormat
<K, V>

MultipleSequenceFile
OutputFormat<K, V>

NullOutputFormat
<K, V>

DBOutputFormat<K, V>

FilterOutputFormat
<K,V>

LazyOutputFormat<K, V>
```

OutputFormat
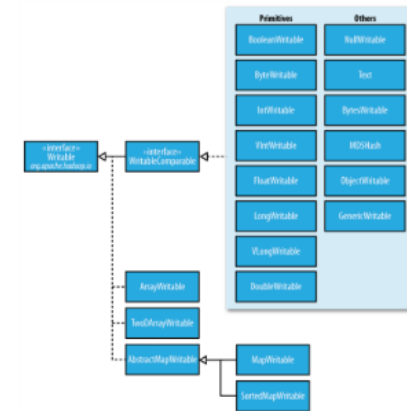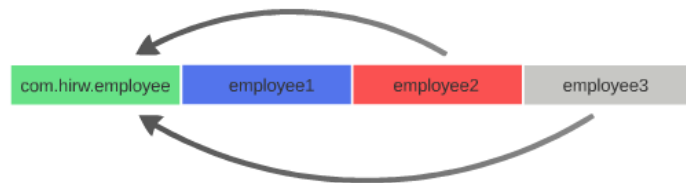
Validate output specifications

RecordWriter implementation to write output files of the job
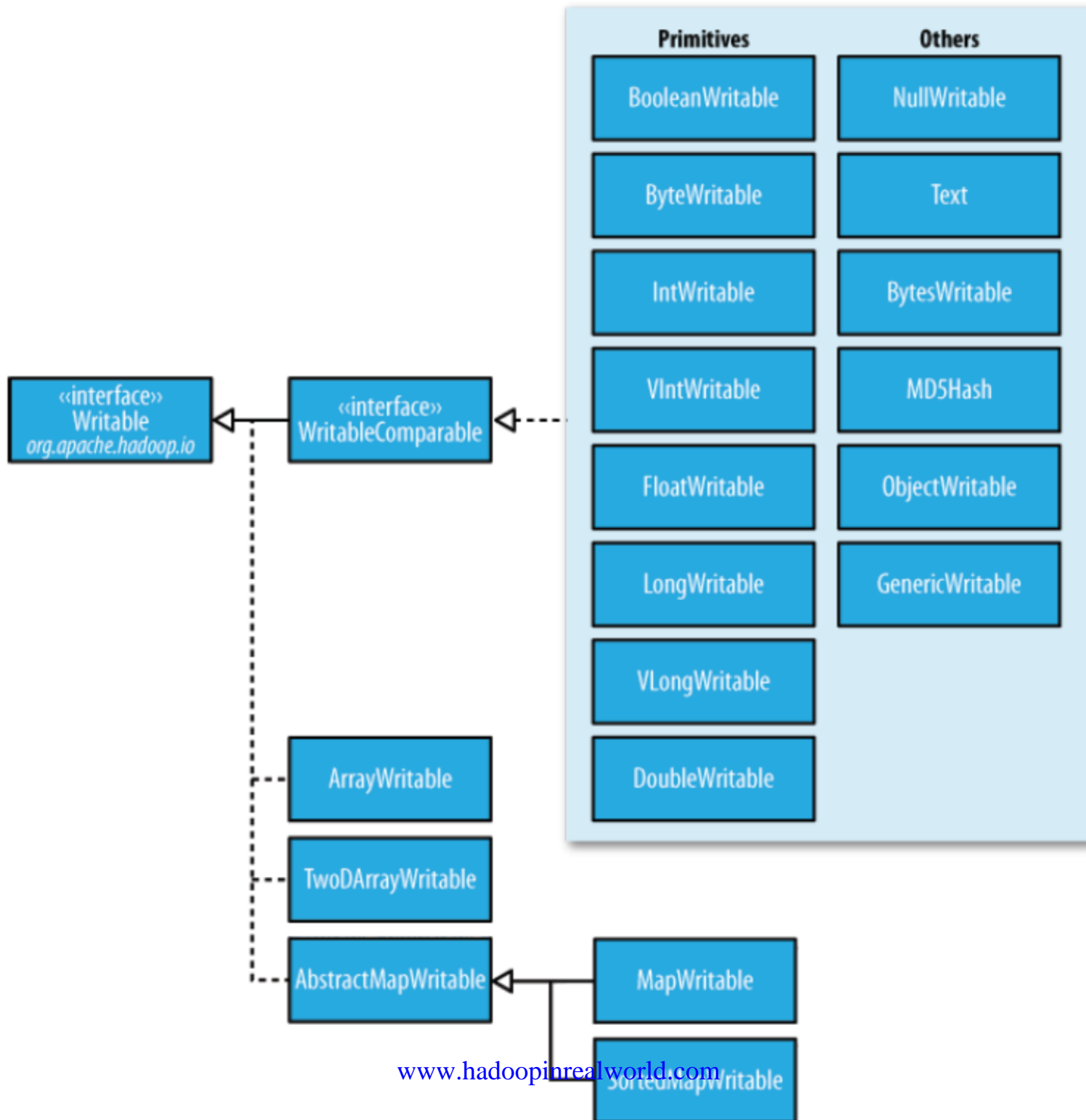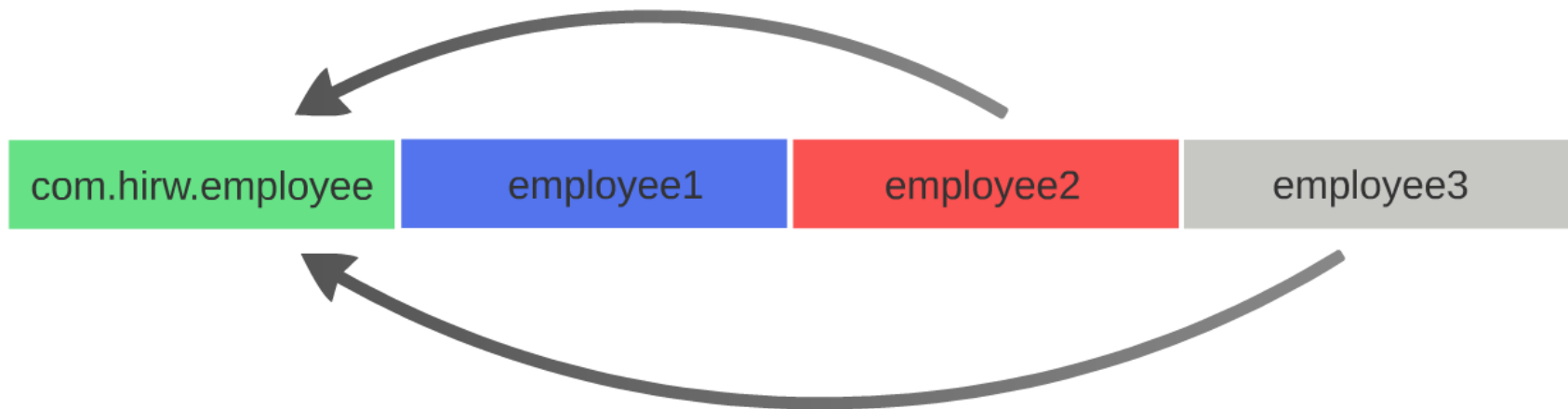
Writable

A serializable object which implements a simple, efficient, serialization protocol

Fast, compact & effective

Primitives	Others
BooleanWritable	NullWritable
ByteWritable	Text
IntWritable	BytesWritable
VIntWritable	MD5Hash
FloatWritable	ObjectWritable
LongWritable	GenericWritable
VLongWritable	
DoubleWritable	

«interface»
Writable
org.apache.hadoop.io

«interface»
WritableComparable

ArrayWritable

TwoDArrayWritable

AbstractMapWritable

MapWritable

SortedMapWritable

www.hadoopinrealworld.com

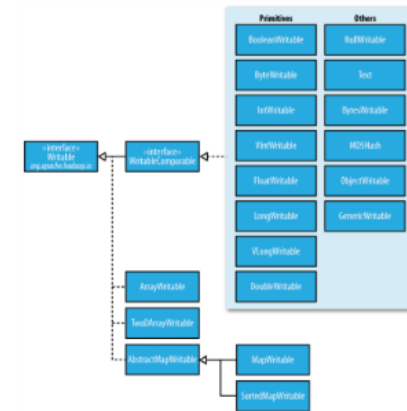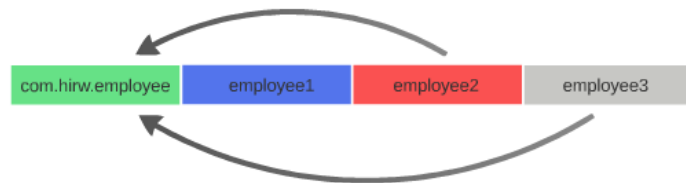| com.hirw.employee | employee1 | employee2 | employee3 |

Writable

A serializable object which implements a simple, efficient, serialization protocol

Fast, compact & effective

MAPPER

- Dataset is divided in to multiple parts - Input Splits

- Each Mapper process an Input Split

- Each Mapper can be called multiple times depending on the content of Input Split

- Mapper will emit Key Value pairs as output

- There will be one or more Mapper in a MapReduce job

Input Split

Mapper

Output
Key - Value

ABC 113
PAM 112
XYW 88

REDUCER

- Reduce function will take Key Value pairs from multiple Map functions as input and Reduce them to output

- Keys are grouped with values. Reduce function is called once per key and its values.

- There could be 0, 1 or more Reduce function for a MapReduce job

Input Split

Mapper

Output Key - Value

ABC 60	ABC 62	XYW 100
PAM 88	XYW 99	JPN 78
XYW 102		

Reducer

ABC 62
JPN 78
XYW 102