Data Processing

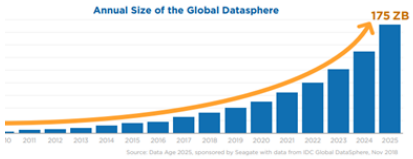
Venkatesh Vinayakarao
venkateshv@cmi.ac.in
http://vvtesh.co.in

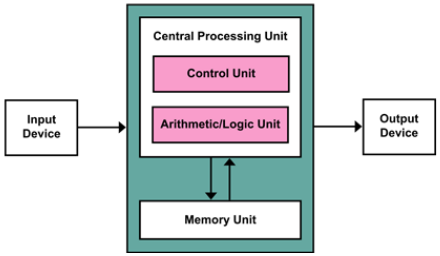
Chennai Mathematical Institute

"In God we trust; all others must bring data". – William Edwards Deming.

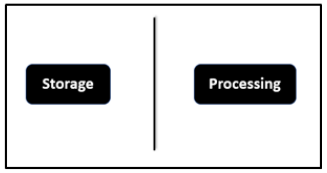
Recap

Name	Size
Byte	8 bits
Kilobyte	1024 bytes
Megabyte	1024 kilobytes
Gigabyte	1024 megabytes
Terabyte	1024 gigabytes
Petabyte	1024 terabytes
Exabyte	1024 petabytes
Zettabyte	1024 exabytes
Yottabyte	1024 zettabytes

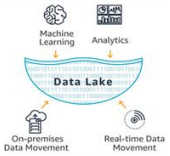




Challenges



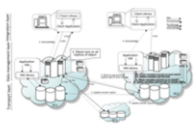
Data Storage



Amazon S3

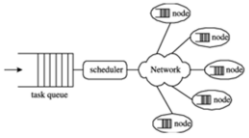
STaaS

Two kinds of Big Data Opportunities

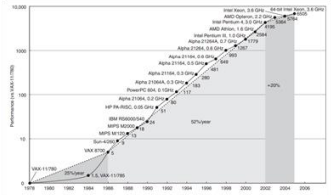


Storage

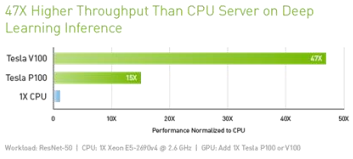
Processing



Data Processing

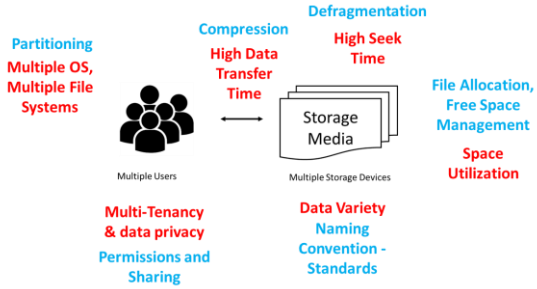


CPU Performance



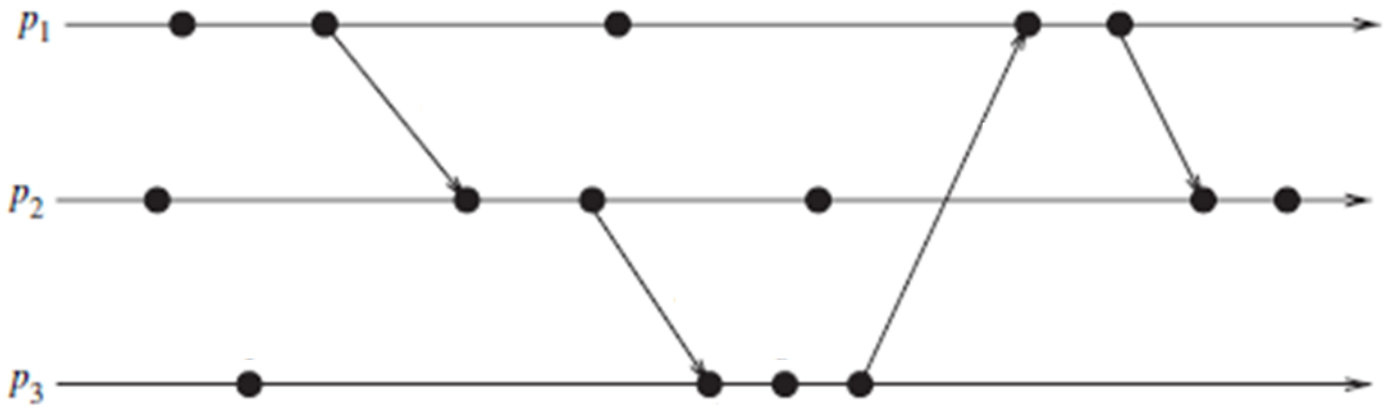
GPU Performance

SuperComputers



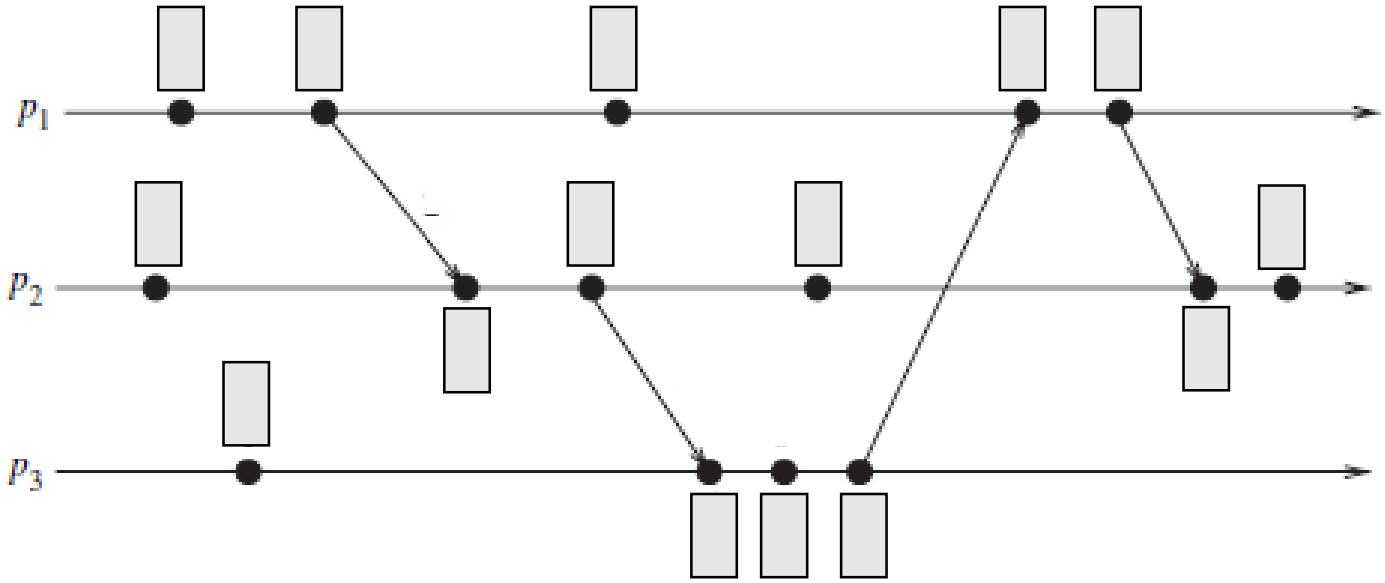
Quiz

- Annotate the following space-time diagram with Lamport's scalar time.



Quiz

- Annotate the following space-time diagram with Lamport's scalar time.

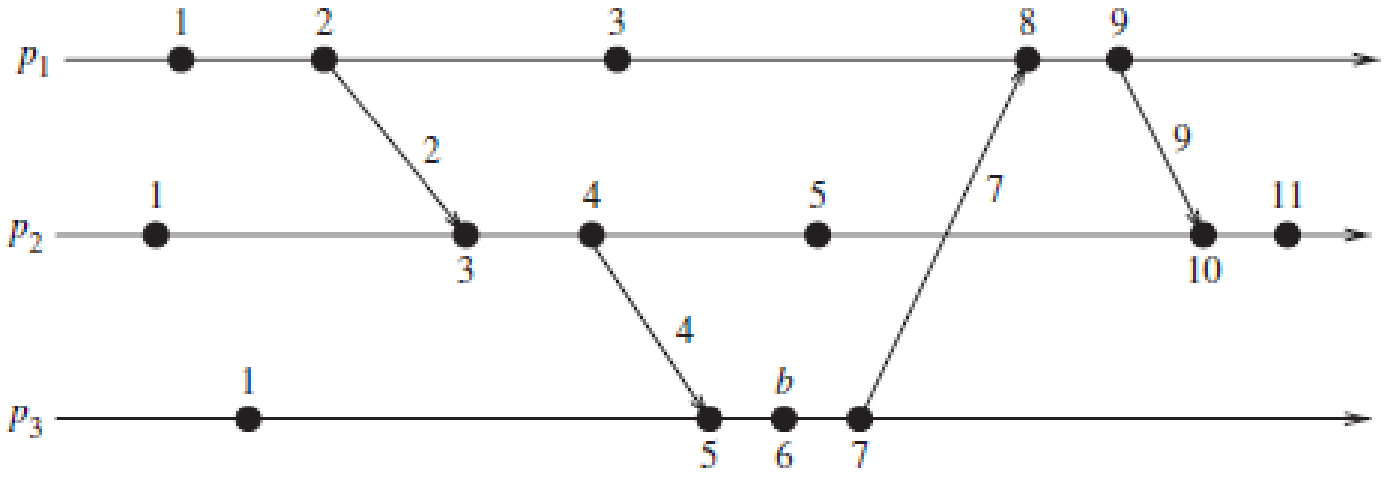


Events of a specific process are in total order.

When messages are passed, $C_i = \max(C_i, C_{msg}) + 1$;

Quiz

- Annotate the following space-time diagram with Lamport's scalar time.



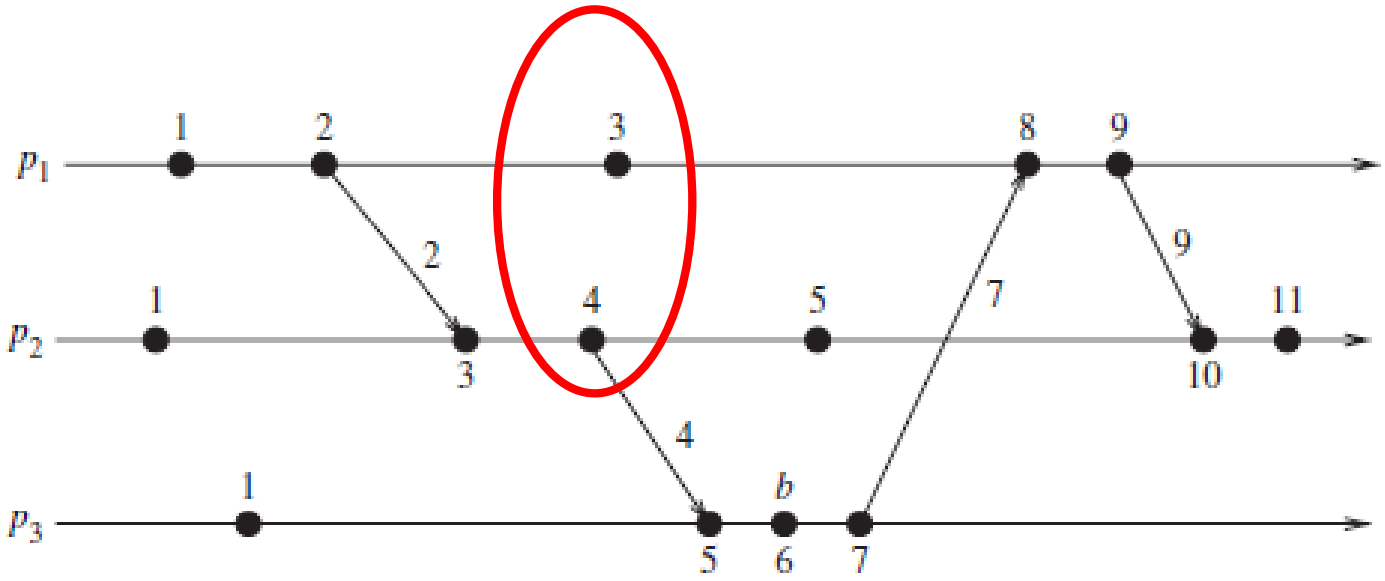
The Problem

- Scalar clocks are consistent
  - e_i happens-before e_j ➔ $C(e_i) < C(e_j)$

- …but not strongly consistent
  - $C(e_i) < C(e_j)$ does not imply e_i happens-before e_j

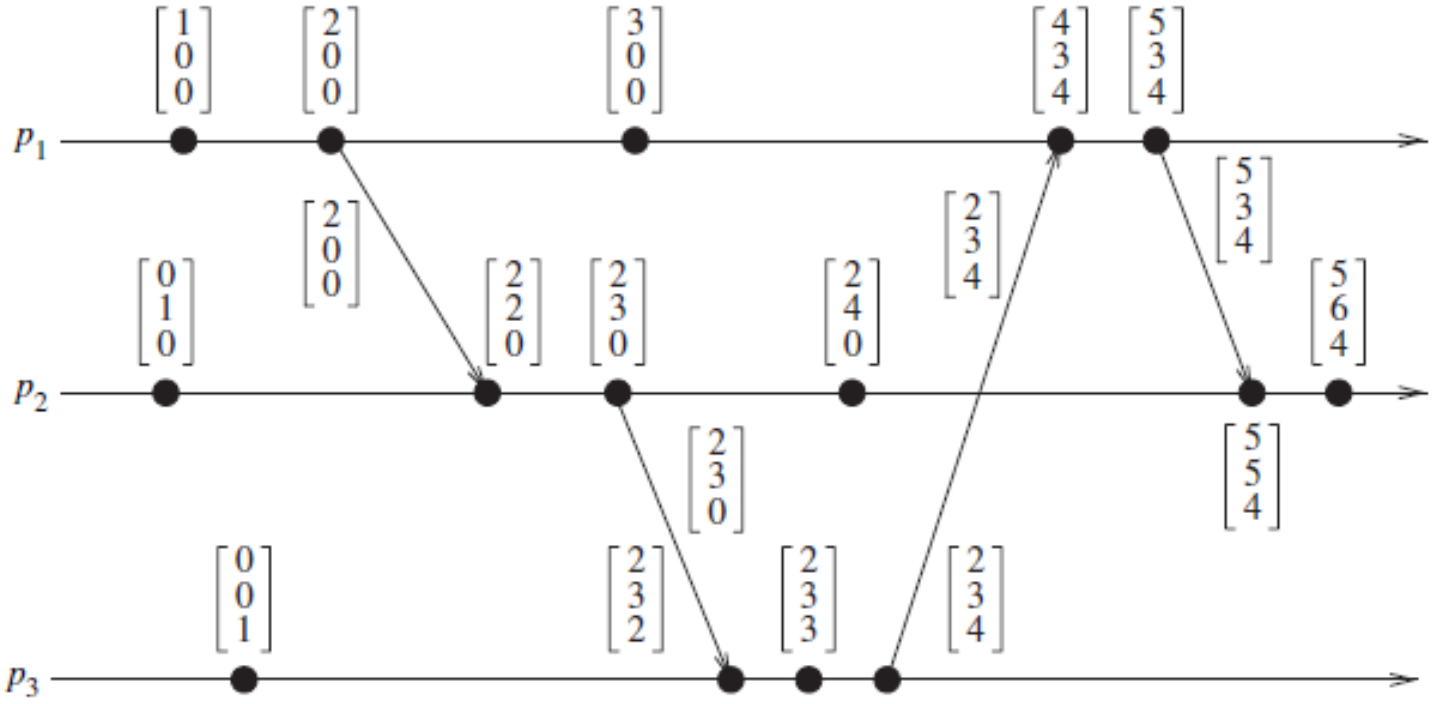
Quiz

- Annotate the following space-time diagram with Lamport's scalar time.



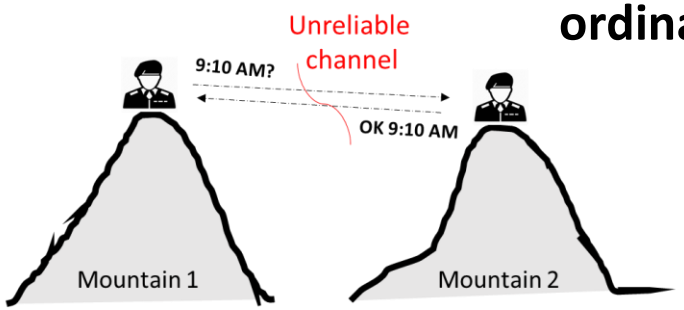
Vector Time

- Annotated space-time diagram with vector time

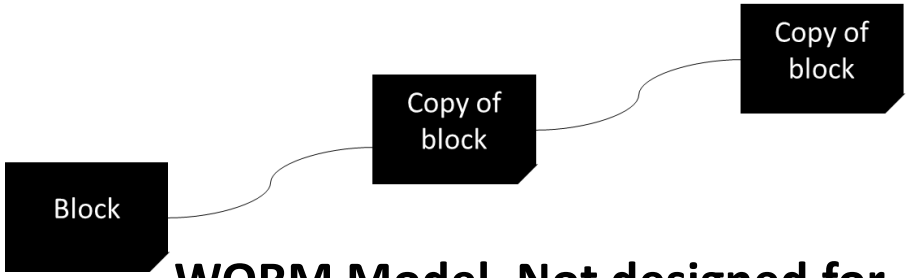


Recap

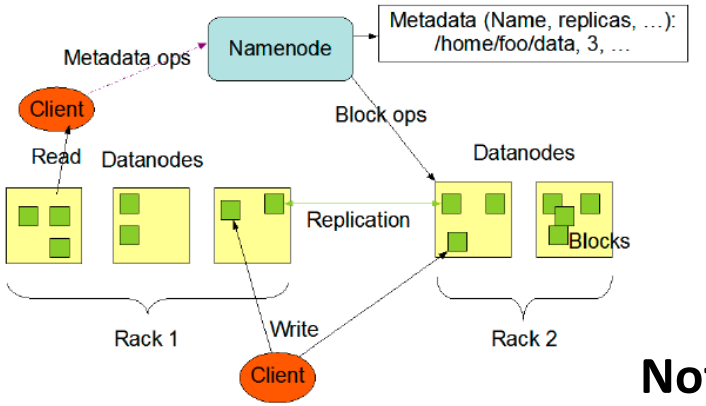
Not designed for co-ordination jobs.



General's Paradox



WORM Model. Not designed for write-many (interactive) jobs.

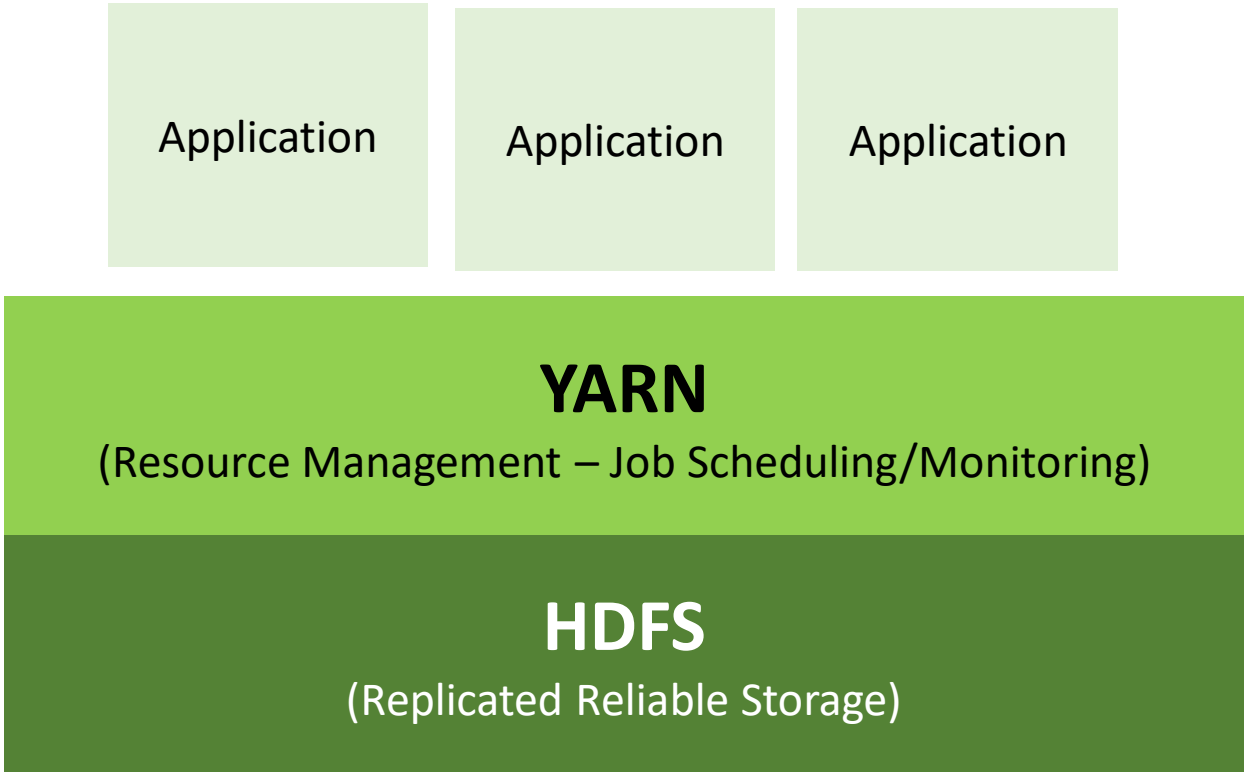


Not designed for small files.

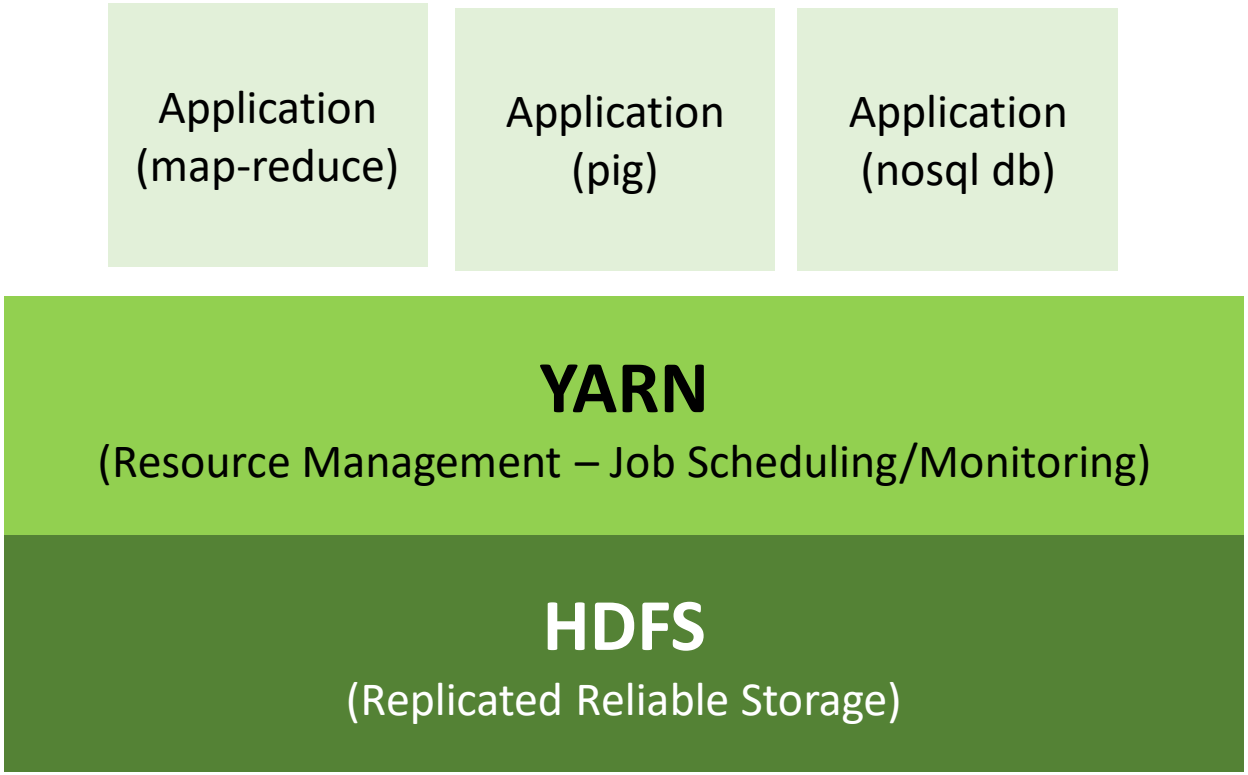
Flex Your Brain!

Distributed Discount Coupon Problem

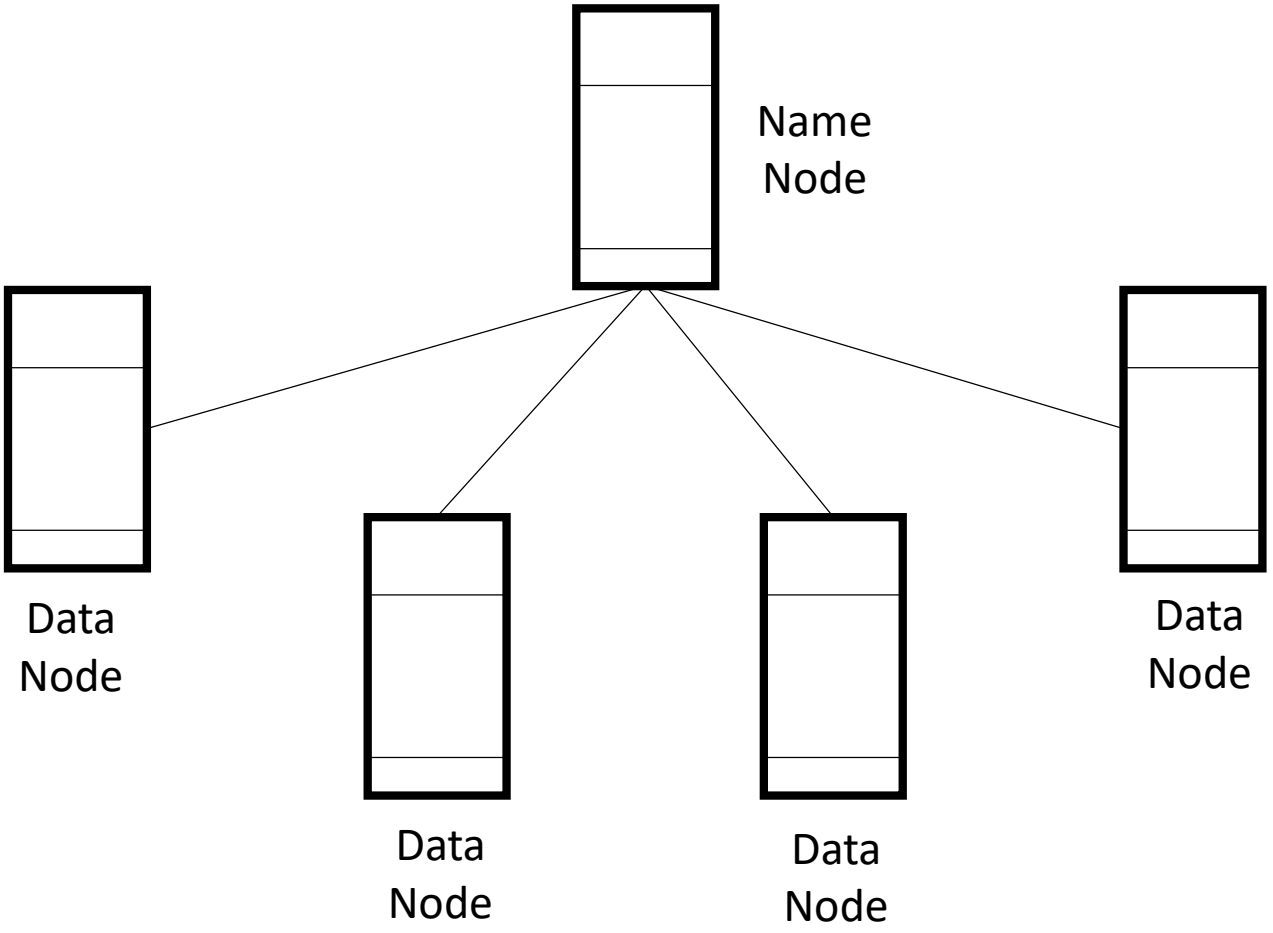
Hadoop Eco System



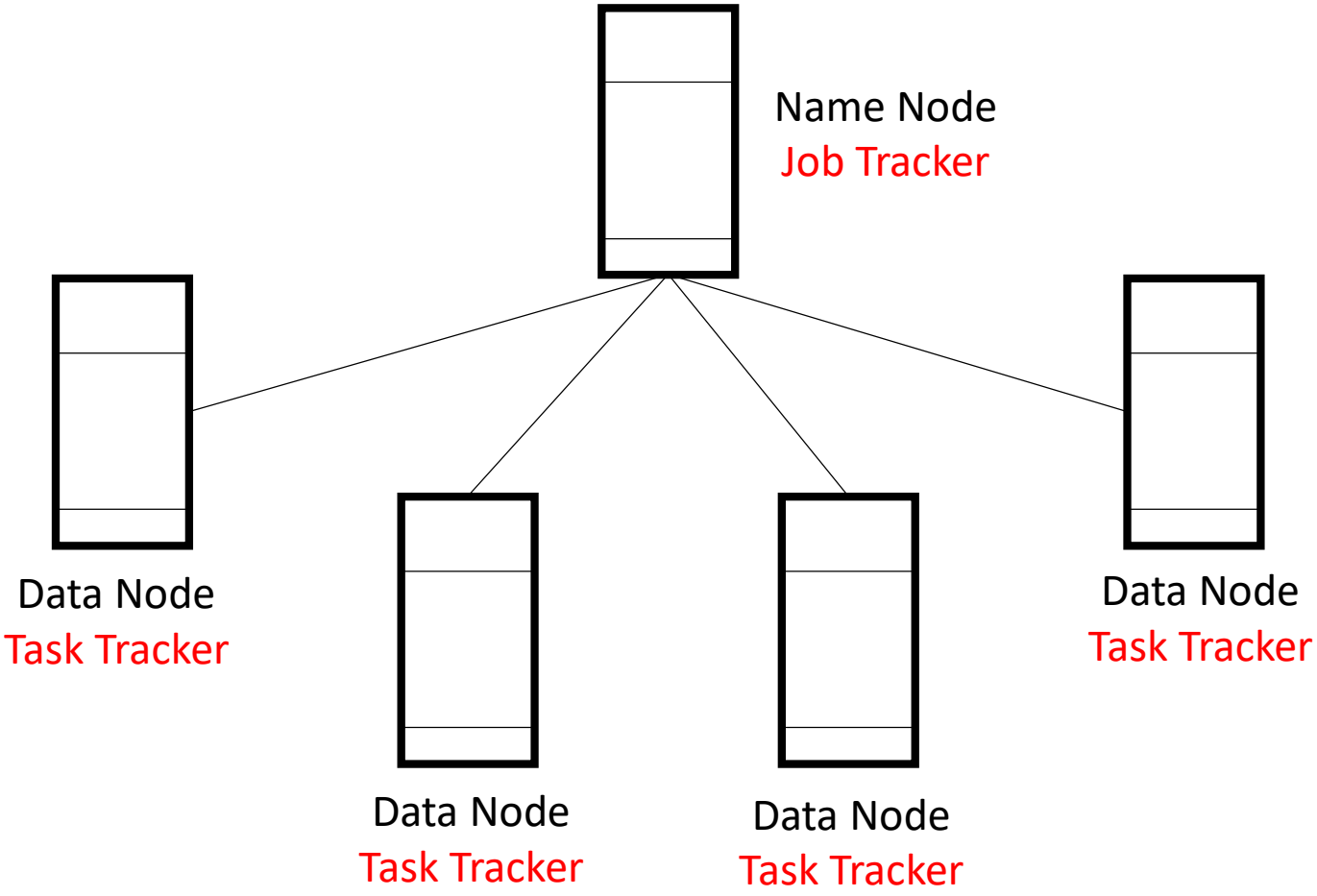
Hadoop Eco System



Hadoop Cluster



Hadoop Cluster

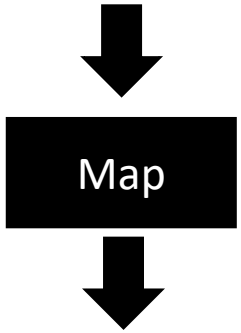


Computing Over the Cluster

“Moving Computation is Cheaper than Moving Data”

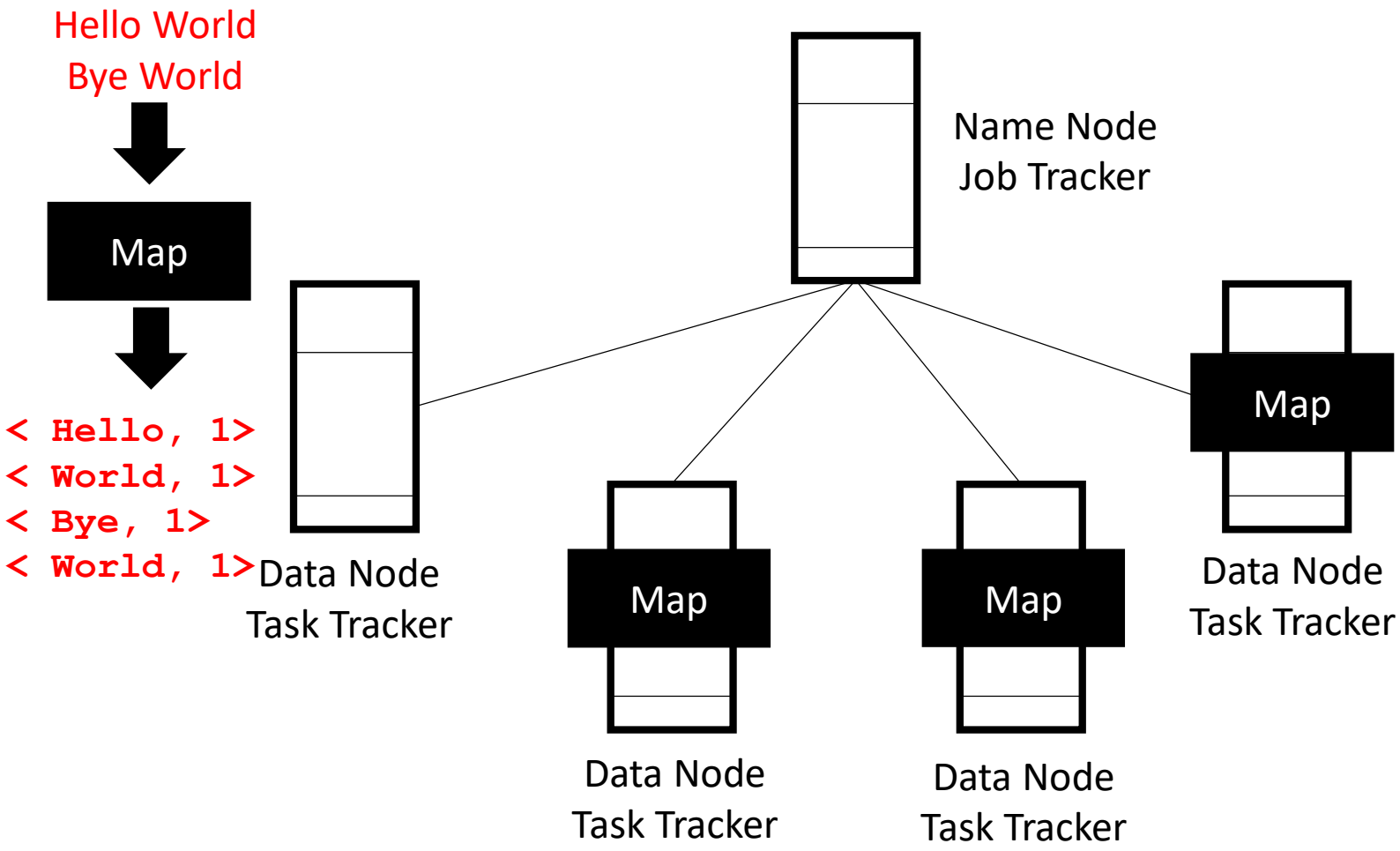
Our Code in a Mapper

Hello World
Bye World

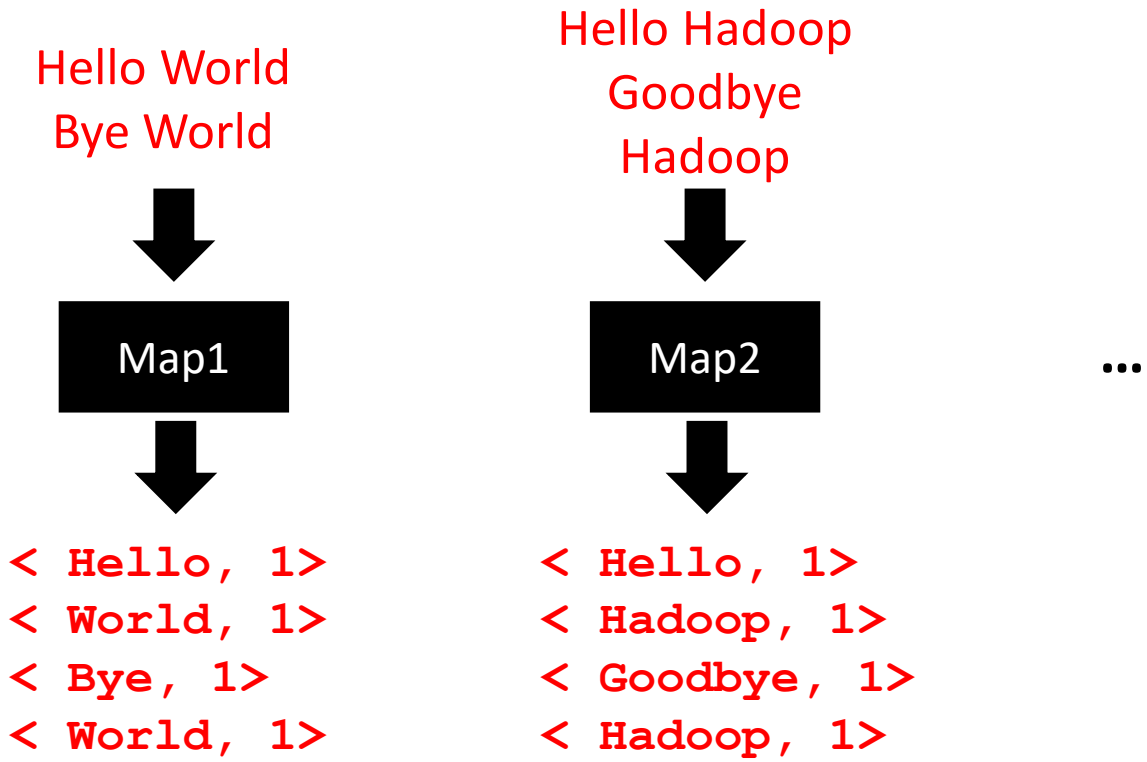



```
< Hello, 1>
< World, 1>
< Bye, 1>
< World, 1>
```

Hadoop Cluster

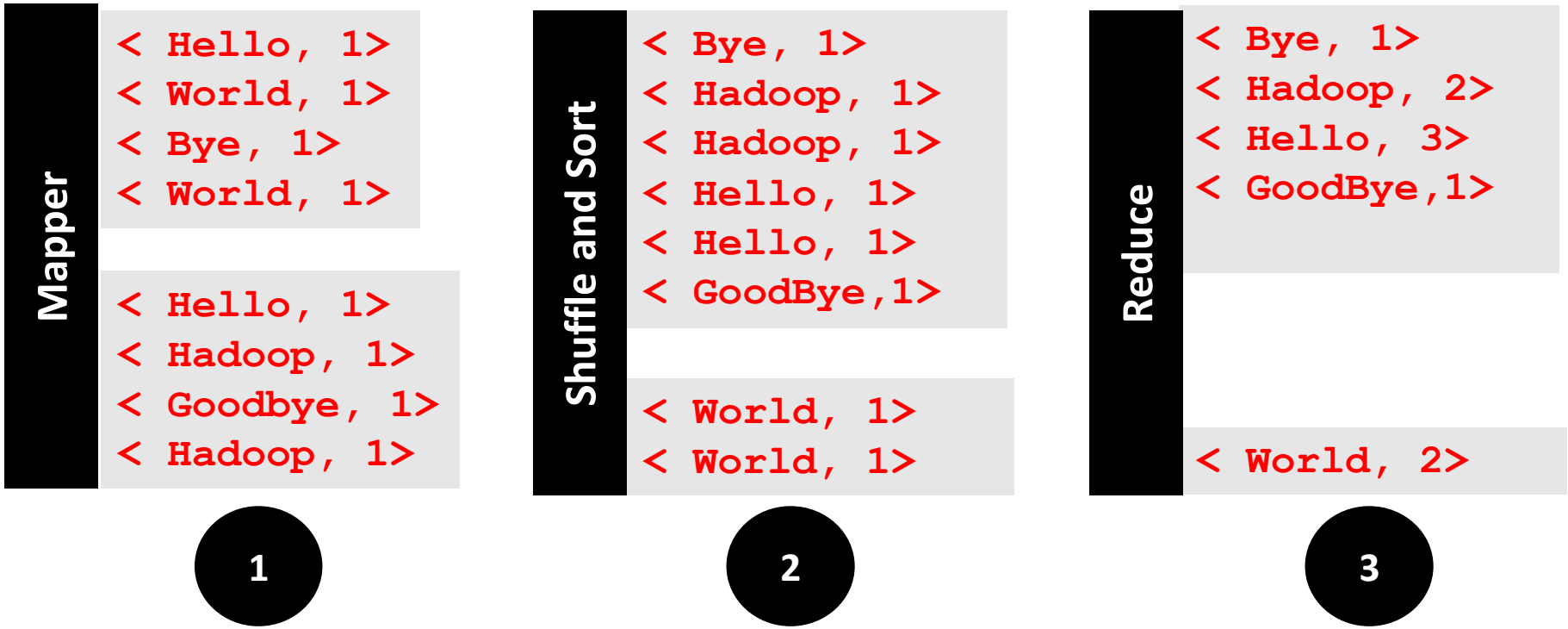


Our Code in a Mapper

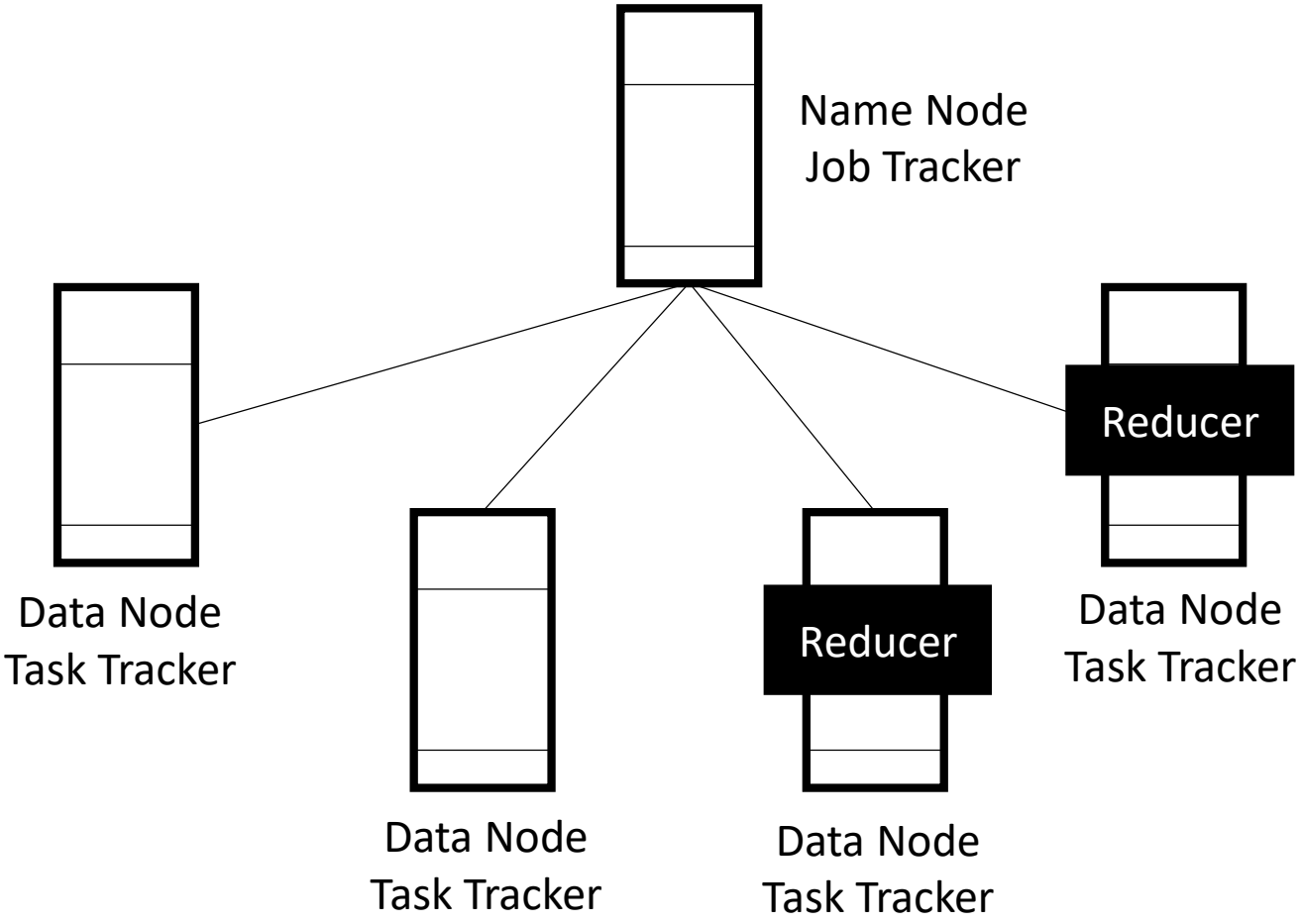


Sorted output of the mappers are associated with reducers.

Map-Reduce Processing

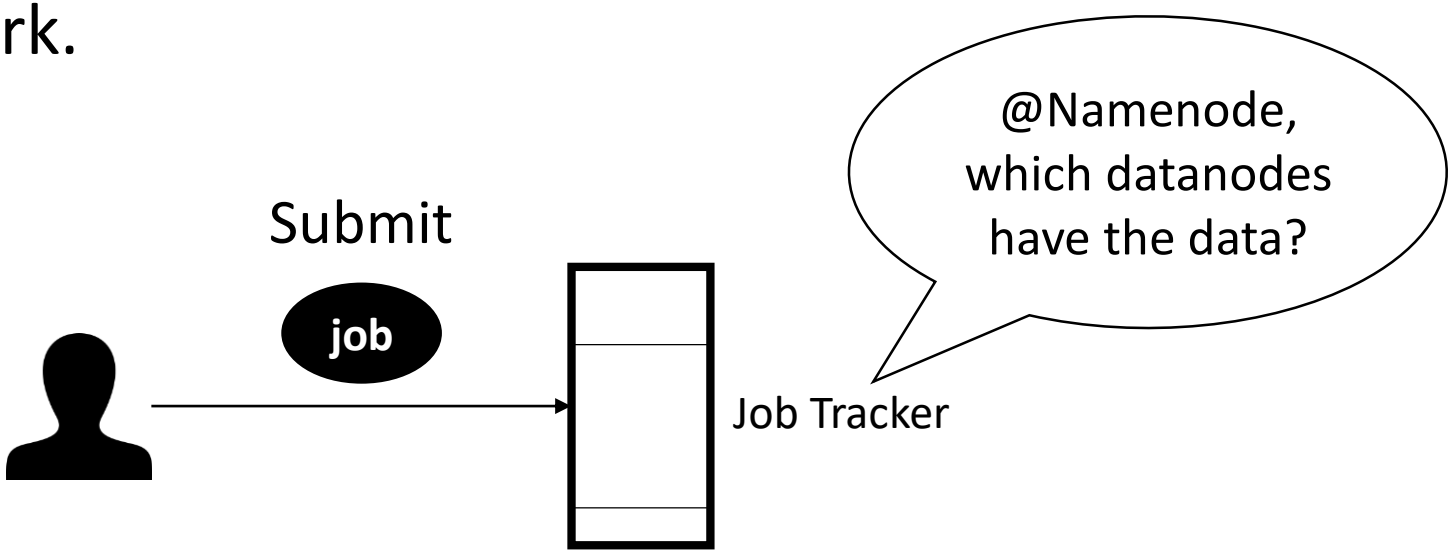


Hadoop Cluster



A Hadoop Map-Reduce Job

- Write the “map” code
- Write the “reduce” code
- Submit the map and reduce code to Hadoop framework.



Submitting a Map-Reduce Job

hadoop jar

/usr/joe/wordcount.jar
org.myorg.WordCount
/usr/joe/wordcount/input
/usr/joe/wordcount/output

Programming Language

- The Hadoop framework itself is mostly written in the Java programming language.
- Hadoop Streaming is a utility that comes with Hadoop.
  - Allows the use of Python for map-reduce coding.

```
hadoop jar
    /usr/lib/hadoop-2.2.0/share/hadoop/tools/lib/hadoop-streaming-2.2.0.jar
    -file /home/cmi/mapper.py
    -mapper mapper.py
    -file   /home/cmi/reducer.py
    -reducer reducer.py
    -input /user/cmi/word
    -output /user/cmi/Wordcount
```

Mapper

```
public void map(Object key, Text value, Context context
                ) throws IOException, InterruptedException {
  StringTokenizer itr = new StringTokenizer(value.toString());
  while (itr.hasMoreTokens()) {
    word.set(itr.nextToken());
    context.write(word, one);
  }
}
```

See https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html for details.

Reducer

```
public void reduce(Text key, Iterable<IntWritable> values,
                   Context context
                   ) throws IOException, InterruptedException {
  int sum = 0;
  for (IntWritable val : values) {
    sum += val.get();
  }
  result.set(sum);
  context.write(key, result);
}
```

Create a Job

```
public static void main(String[] args) throws Exception {
  Configuration conf = new Configuration();
  Job job = Job.getInstance(conf, "word count");
  job.setJarByClass(WordCount.class);
  job.setMapperClass(TokenizerMapper.class);
  job.setCombinerClass(IntSumReducer.class);
  job.setReducerClass(IntSumReducer.class);
  job.setOutputKeyClass(Text.class);
  job.setOutputValueClass(IntWritable.class);
  FileInputFormat.addInputPath(job, new Path(args[0]));
  FileOutputFormat.setOutputPath(job, new Path(args[1]));
  System.exit(job.waitForCompletion(true) ? 0 : 1);
}
```

Submit Job to Hadoop

```
$ bin/hadoop jar wc.jar WordCount /user/joe/wordcount/input
                  /user/joe/wordcount/output
```

```
$ bin/hadoop fs -ls /user/joe/wordcount/input/
/user/joe/wordcount/input/file01
/user/joe/wordcount/input/file02

$ bin/hadoop fs -cat /user/joe/wordcount/input/file01
Hello World Bye World

$ bin/hadoop fs -cat /user/joe/wordcount/input/file02
Hello Hadoop Goodbye Hadoop
```

Output

```
$ bin/hadoop fs -cat /user/joe/wordcount/output/part-r-00000
Bye 1
Goodbye 1
Hadoop 2
Hello 2
World 2
```

Key and Value

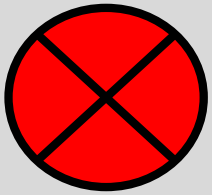
Together, they make a lovely pair!

Role of Key-Value Pairs

- Map-Reduce coding
- Key-Value datastores (Amazon DynamoDB)
- Windows registry entries
- JSON files
- Hashing
- …

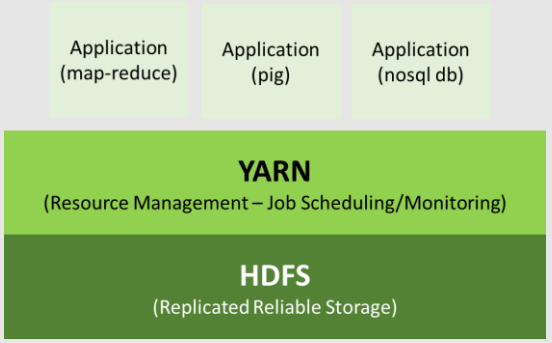
Summary

When not to use Hadoop?



No Interactive Jobs
No Jobs Requiring Co-ordination
No Small Files

Hadoop Architecture



Map-reduce Model