Introduction to Big Data and Hadoop

Module 1 – Big Data Fundamentals

Godson D'silva – Research Scholar

Big Data Fundamentals

- What is Big Data
- Hadoop Ecosystem
- HDFS Concepts
- MapReduce Concepts

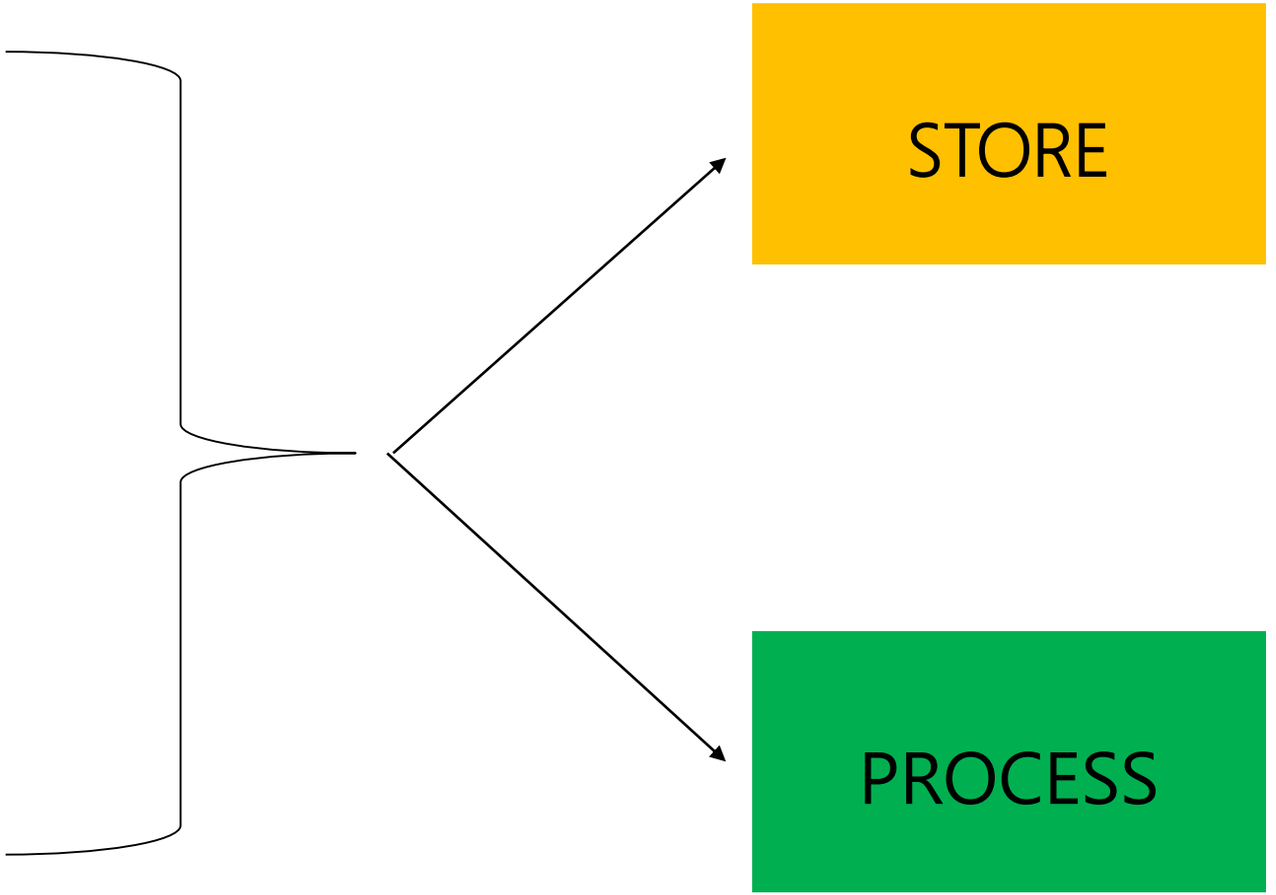
Data Sources :

> Phone Data

> Online Stores

> Medicines

> Research ……etc



What is Big Data ?

Definition :

Its Data that's too Big to be Process on a Single Machine

Challenges in Big Data

Data volumes

Volume refers to the Size of the Data your Dealing with …

Data variety

Variety refers to fact that the data is often coming from different data sources

Data velocity

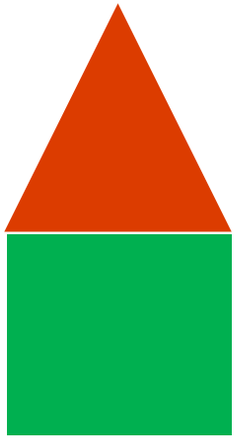
Velocity refers to the speed at which the data is generated.

Data Volume

$100,000+
1980

Price/GB

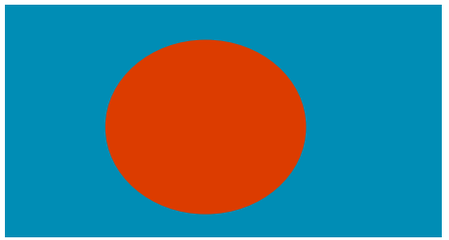
$ 0.10+
2016



$0.10
Home

SAN

- Store & process



$$$ Reliable
Storage

Data Volume :

- We need a *cheaper way to store reliable*, also needs to read and process data efficiently.

- Storing a Data on SAN is not that hard but *Streaming the data on the SAN across the network* to a central processor can take a long time.

Data Variety :

- People have used databases such as SQL Server, MySQL, or Big data warehouse from companies like Oracle, IBM to Store there Data.

- The problem is to store data in system like that the data needs to be fit in a pre-defined tables.

- And lot of data we deal with nowadays tends to be *Un-Structured Or Semi-Structured Data.*

Data Variety :

- Structured Data : Eg Relational Tables-

empID	empName	empSalary	empAddress
101	Ramsey	$7800054	wales
102	Ozil	$9850000000	germany

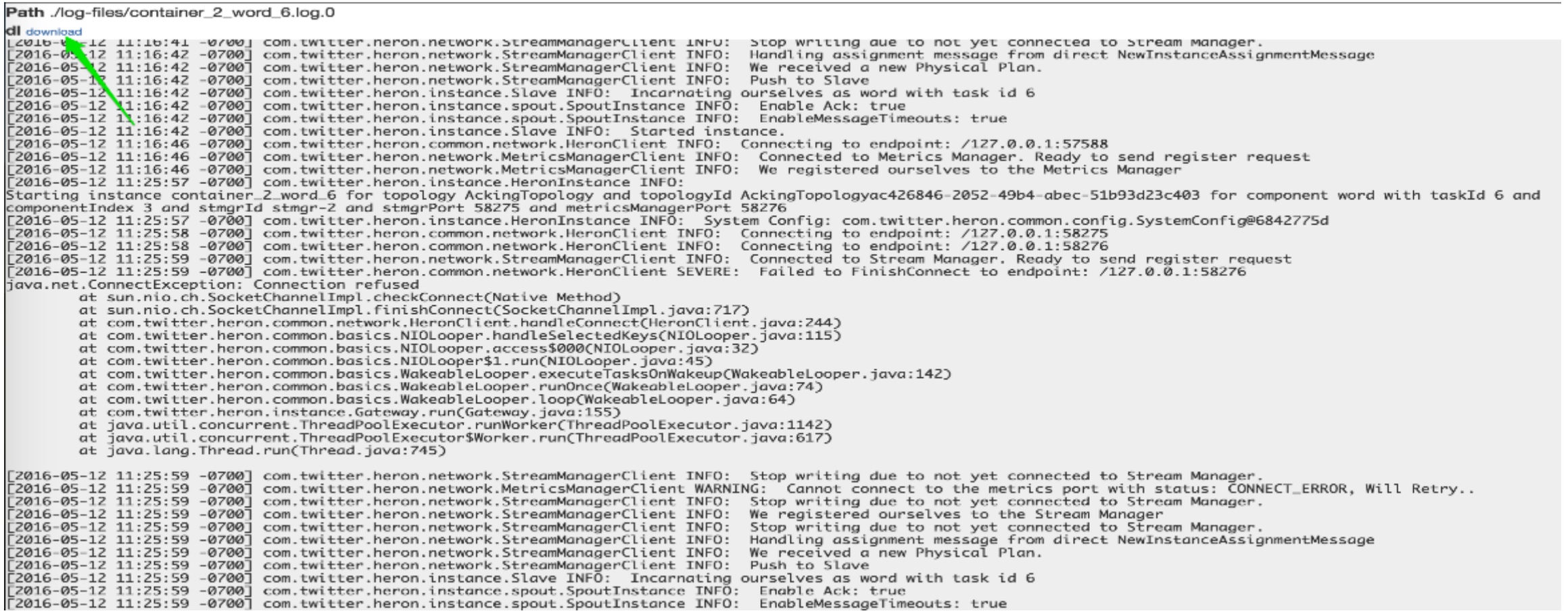
Data Variety :

- Semi-Structured Data : Eg XML files

```
<note>
<to>Tove</to>
<from>Jani</from>
<heading>Reminder</heading>
<body>Don't forget me this weekend!</body>
</note>
```

Data Variety :

- Un-Structured Data : Eg Logs



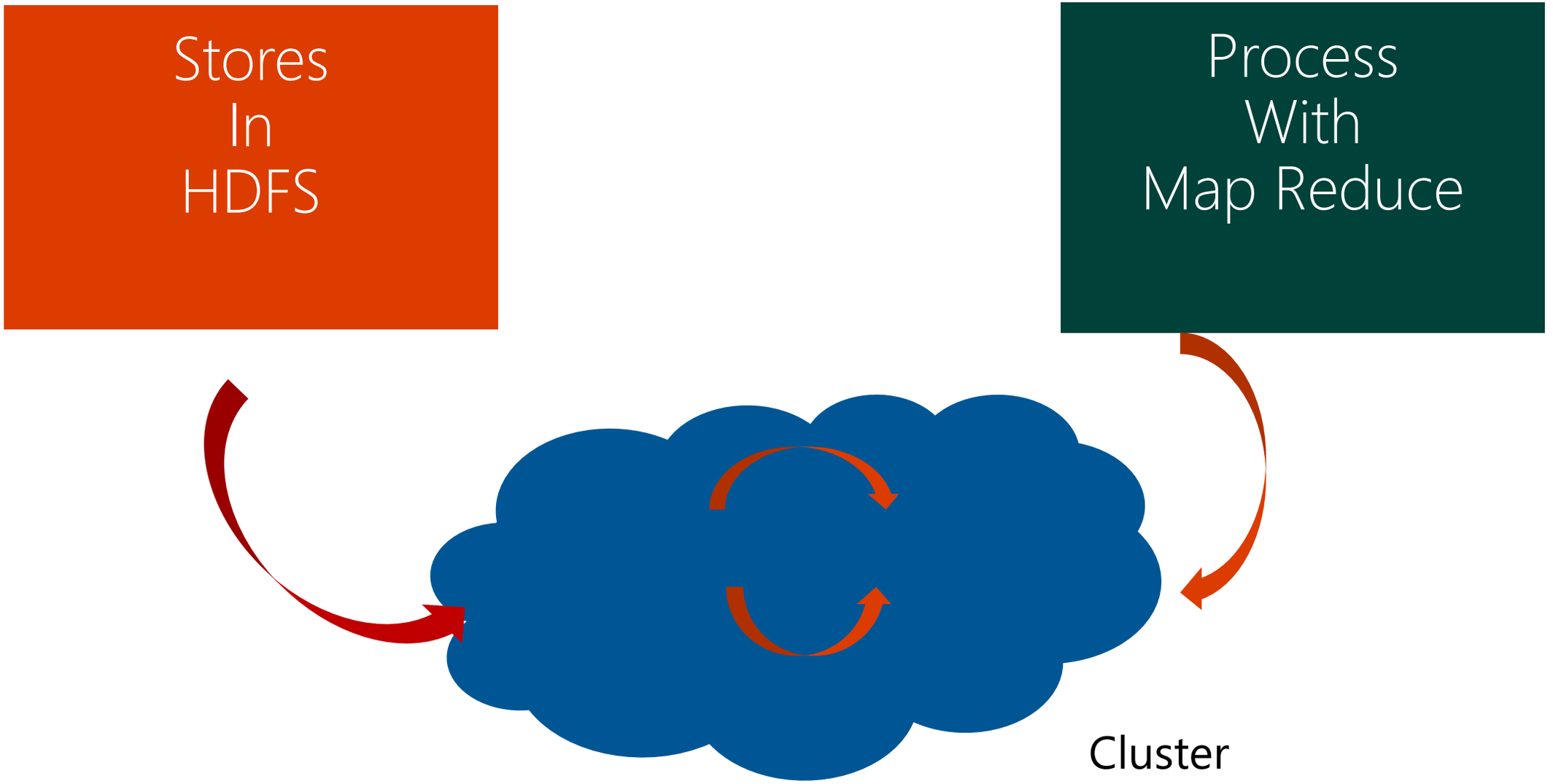
```
Path ./log-files/container_2_word_6.log.0
dl download
[2016-05-12 11:16:41 -0700] com.twitter.heron.network.StreamManagerClient INFO:  Stop writing due to not yet connected to Stream Manager.
[2016-05-12 11:16:42 -0700] com.twitter.heron.network.StreamManagerClient INFO:  Handling assignment message from direct NewInstanceAssignmentMessage
[2016-05-12 11:16:42 -0700] com.twitter.heron.network.StreamManagerClient INFO:  We received a new Physical Plan.
[2016-05-12 11:16:42 -0700] com.twitter.heron.network.StreamManagerClient INFO:  Push to Slave
[2016-05-12 11:16:42 -0700] com.twitter.heron.instance.Slave INFO:  Incarnating ourselves as word with task id 6
[2016-05-12 11:16:42 -0700] com.twitter.heron.instance.spout.SpoutInstance INFO:  Enable Ack: true
[2016-05-12 11:16:42 -0700] com.twitter.heron.instance.spout.SpoutInstance INFO:  EnableMessageTimeouts: true
[2016-05-12 11:16:42 -0700] com.twitter.heron.instance.Slave INFO:  Started instance.
[2016-05-12 11:16:46 -0700] com.twitter.heron.common.network.HeronClient INFO:  Connecting to endpoint: /127.0.0.1:57588
[2016-05-12 11:16:46 -0700] com.twitter.heron.network.MetricsManagerClient INFO:  Connected to Metrics Manager. Ready to send register request
[2016-05-12 11:16:46 -0700] com.twitter.heron.network.MetricsManagerClient INFO:  We registered ourselves to the Metrics Manager
[2016-05-12 11:25:57 -0700] com.twitter.heron.instance.HeronInstance INFO:
Starting instance container_2_word_6 for topology AckingTopology and topologyId AckingTopologyac426846-2052-49b4-abec-51b93d23c403 for component word with taskId 6 and
componentIndex 3 and stmgrId stmgr-2 and stmgrPort 58275 and metricsManagerPort 58276
[2016-05-12 11:25:57 -0700] com.twitter.heron.instance.HeronInstance INFO:  System Config: com.twitter.heron.common.config.SystemConfig@6842775d
[2016-05-12 11:25:58 -0700] com.twitter.heron.common.network.HeronClient INFO:  Connecting to endpoint: /127.0.0.1:58275
[2016-05-12 11:25:58 -0700] com.twitter.heron.common.network.HeronClient INFO:  Connecting to endpoint: /127.0.0.1:58276
[2016-05-12 11:25:59 -0700] com.twitter.heron.network.StreamManagerClient INFO:  Connected to Stream Manager. Ready to send register request
[2016-05-12 11:25:59 -0700] com.twitter.heron.common.network.HeronClient SEVERE:  Failed to FinishConnect to endpoint: /127.0.0.1:58276
java.net.ConnectException: Connection refused
        at sun.nio.ch.SocketChannelImpl.checkConnect(Native Method)
        at sun.nio.ch.SocketChannelImpl.finishConnect(SocketChannelImpl.java:717)
        at com.twitter.heron.common.network.HeronClient.handleConnect(HeronClient.java:244)
        at com.twitter.heron.common.basics.NIOLooper.handleSelectedKeys(NIOLooper.java:115)
        at com.twitter.heron.common.basics.NIOLooper.access$000(NIOLooper.java:32)
        at com.twitter.heron.common.basics.NIOLooper$1.run(NIOLooper.java:45)
        at com.twitter.heron.common.basics.WakeableLooper.executeTasksOnWakeup(WakeableLooper.java:142)
        at com.twitter.heron.common.basics.WakeableLooper.runOnce(WakeableLooper.java:74)
        at com.twitter.heron.common.basics.WakeableLooper.loop(WakeableLooper.java:64)
        at com.twitter.heron.instance.Gateway.run(Gateway.java:155)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
        at java.lang.Thread.run(Thread.java:745)

[2016-05-12 11:25:59 -0700] com.twitter.heron.network.StreamManagerClient INFO:  Stop writing due to not yet connected to Stream Manager.
[2016-05-12 11:25:59 -0700] com.twitter.heron.network.MetricsManagerClient WARNING:  Cannot connect to the metrics port with status: CONNECT_ERROR, Will Retry..
[2016-05-12 11:25:59 -0700] com.twitter.heron.network.StreamManagerClient INFO:  Stop writing due to not yet connected to Stream Manager.
[2016-05-12 11:25:59 -0700] com.twitter.heron.network.StreamManagerClient INFO:  We registered ourselves to the Stream Manager
[2016-05-12 11:25:59 -0700] com.twitter.heron.network.StreamManagerClient INFO:  Stop writing due to not yet connected to Stream Manager.
[2016-05-12 11:25:59 -0700] com.twitter.heron.network.StreamManagerClient INFO:  Handling assignment message from direct NewInstanceAssignmentMessage
[2016-05-12 11:25:59 -0700] com.twitter.heron.network.StreamManagerClient INFO:  We received a new Physical Plan.
[2016-05-12 11:25:59 -0700] com.twitter.heron.network.StreamManagerClient INFO:  Push to Slave
[2016-05-12 11:25:59 -0700] com.twitter.heron.instance.Slave INFO:  Incarnating ourselves as word with task id 6
[2016-05-12 11:25:59 -0700] com.twitter.heron.instance.spout.SpoutInstance INFO:  Enable Ack: true
[2016-05-12 11:25:59 -0700] com.twitter.heron.instance.spout.SpoutInstance INFO:  EnableMessageTimeouts: true
```

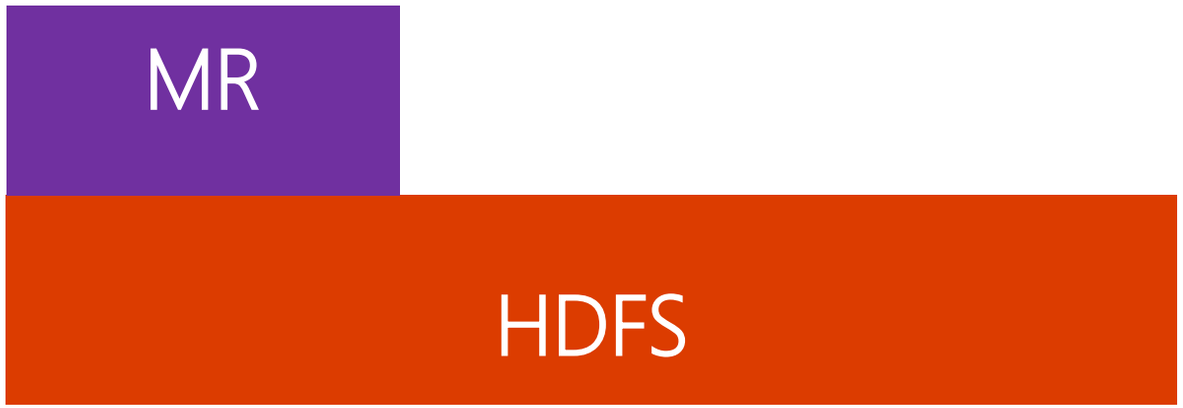
Data Velocity:

- Speed at which the data arrives, ready to be process.
- We need to accept and store the data even when its coming at rate of TB/Day.
- If we cant store as it arrives, we end up discarding some of it and that's what we don't want.
- Eg : Product Recommendations..

Hadoop

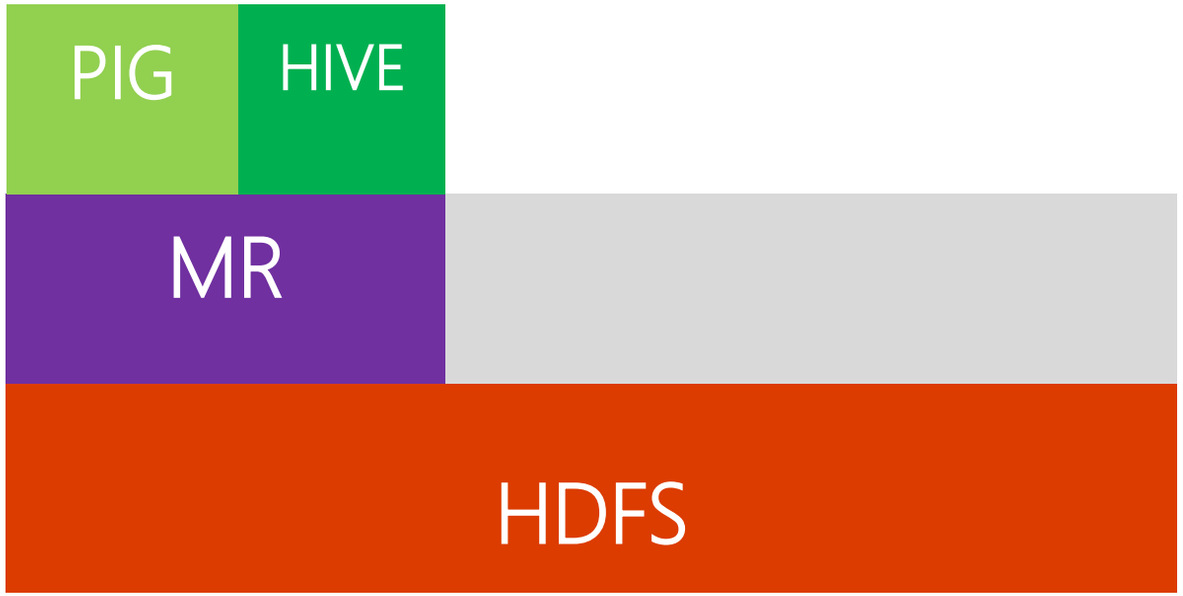


Hadoop Ecosystem

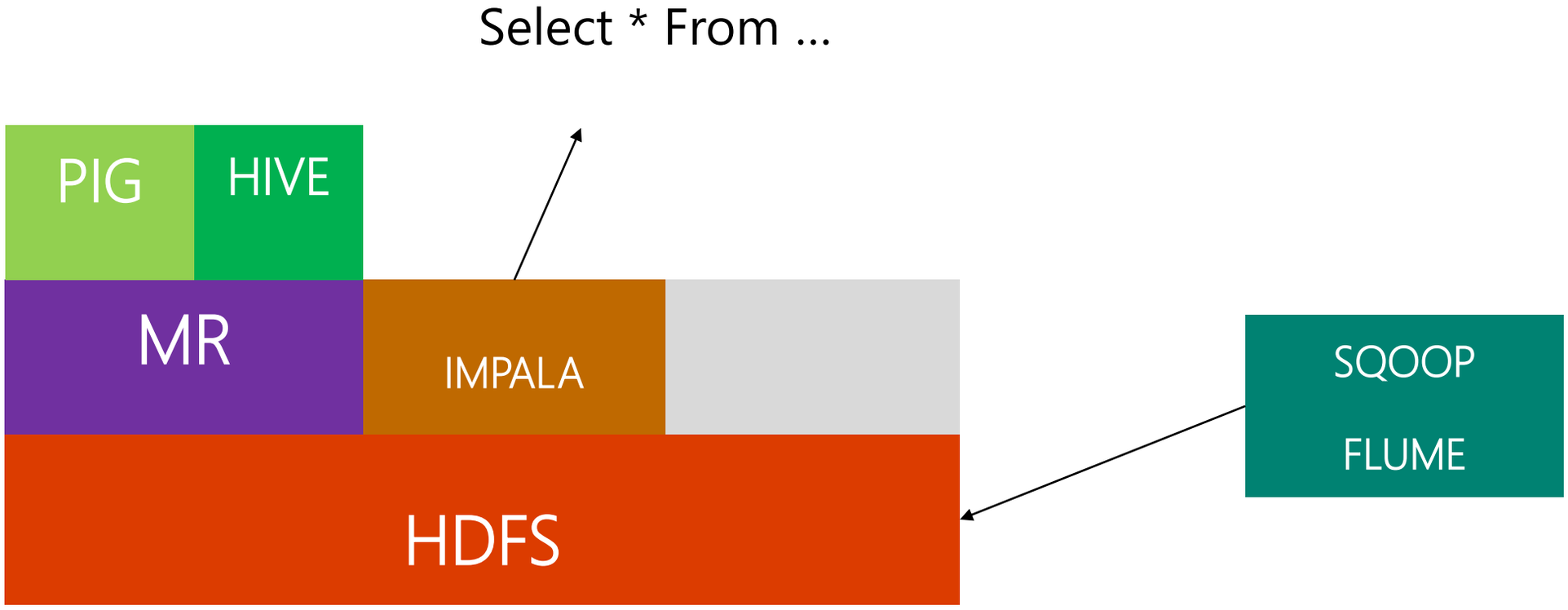


Hadoop Ecosystem

Select * From …

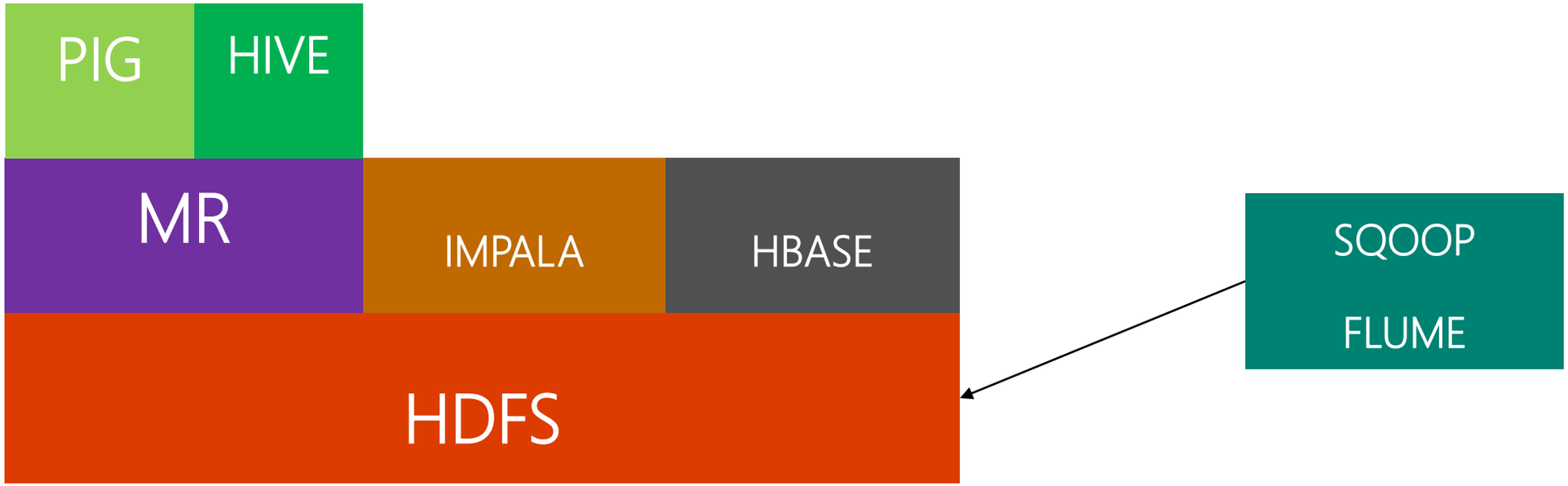


Hadoop Ecosystem



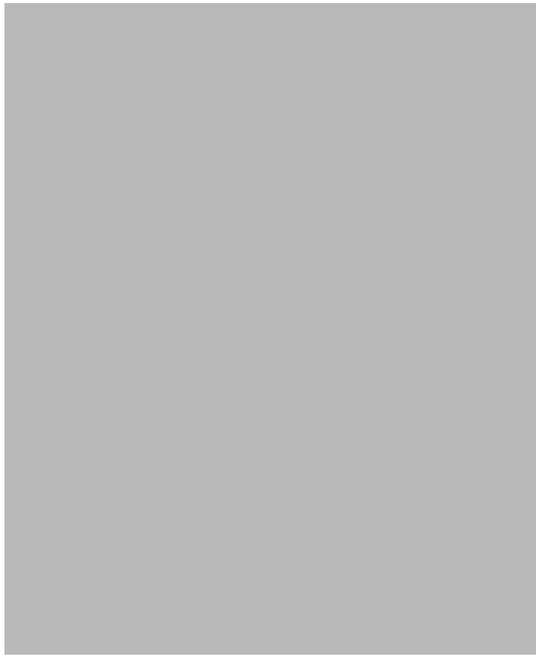
Hadoop Ecosystem

HUE, OOZIE, MAHOUT …



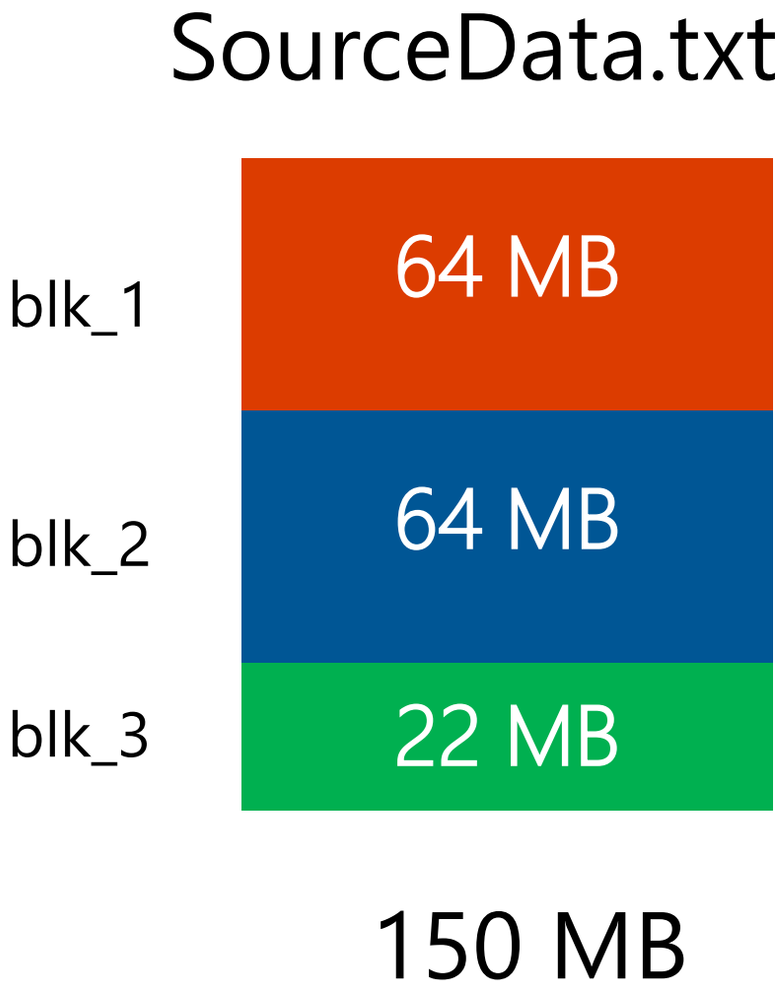
HDFS :

SourceData.txt

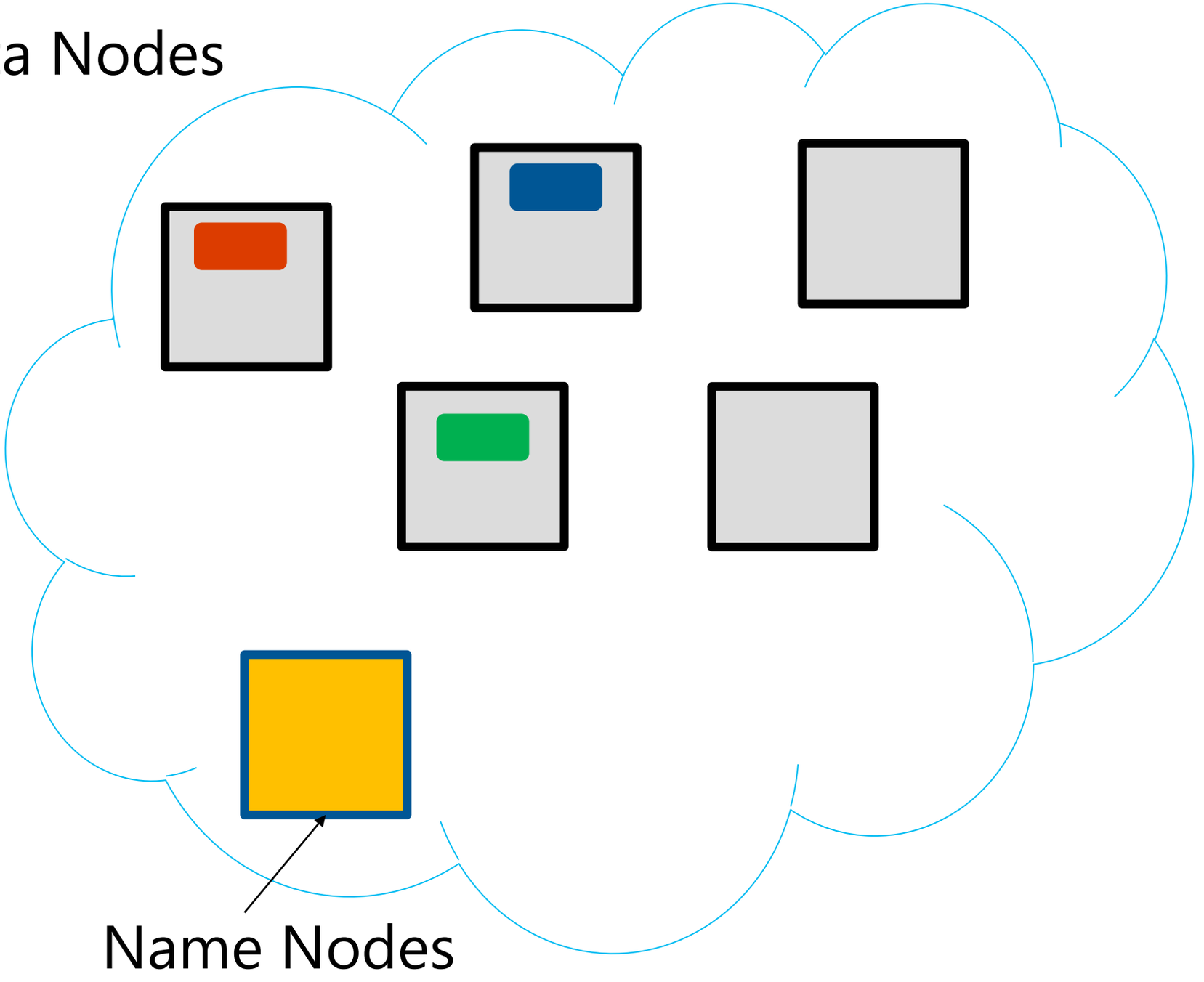


150 MB

HDFS :



HDFS :

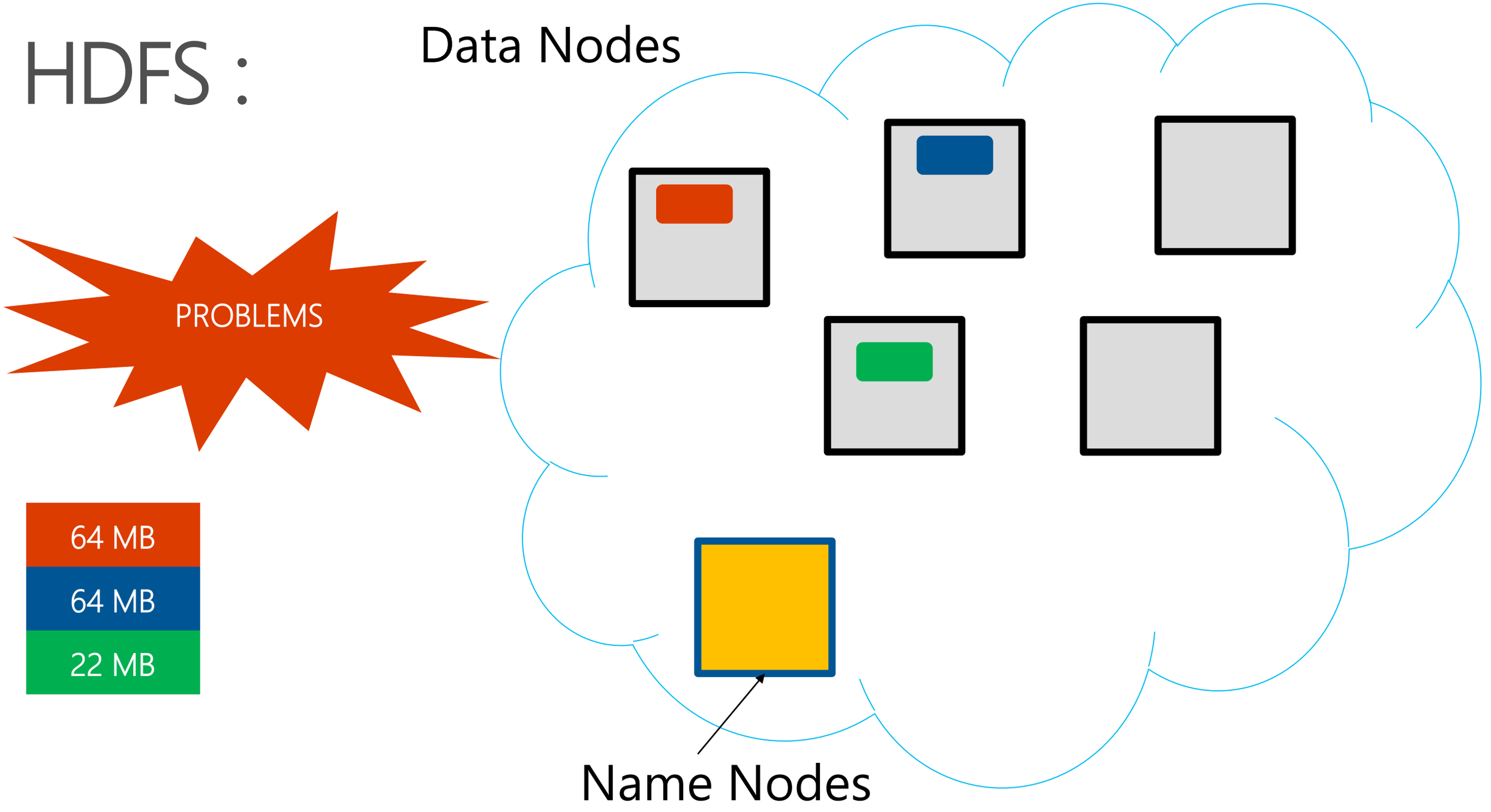


64 MB

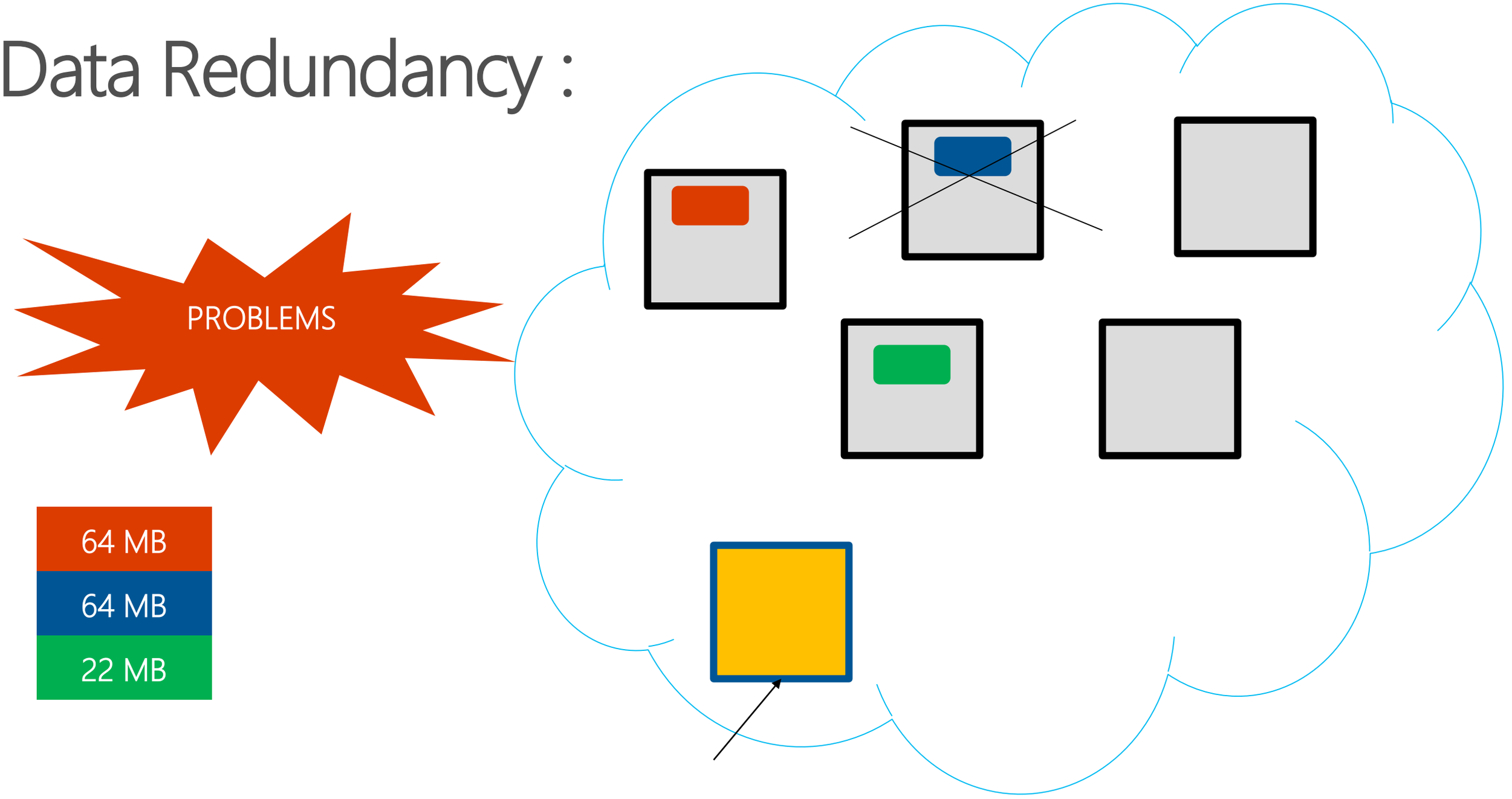
64 MB

22 MB

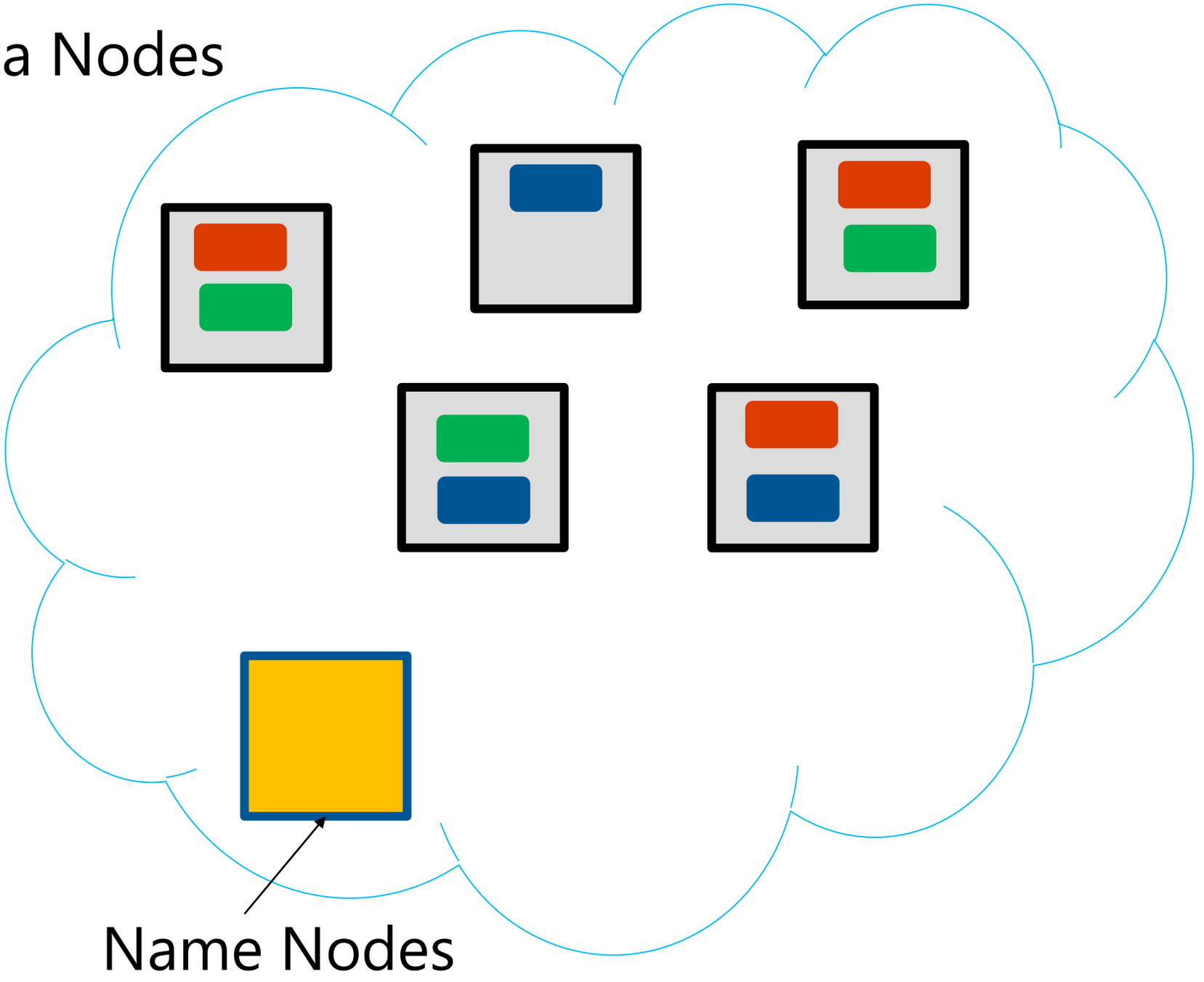
HDFS :



Data Redundancy :



Solution :

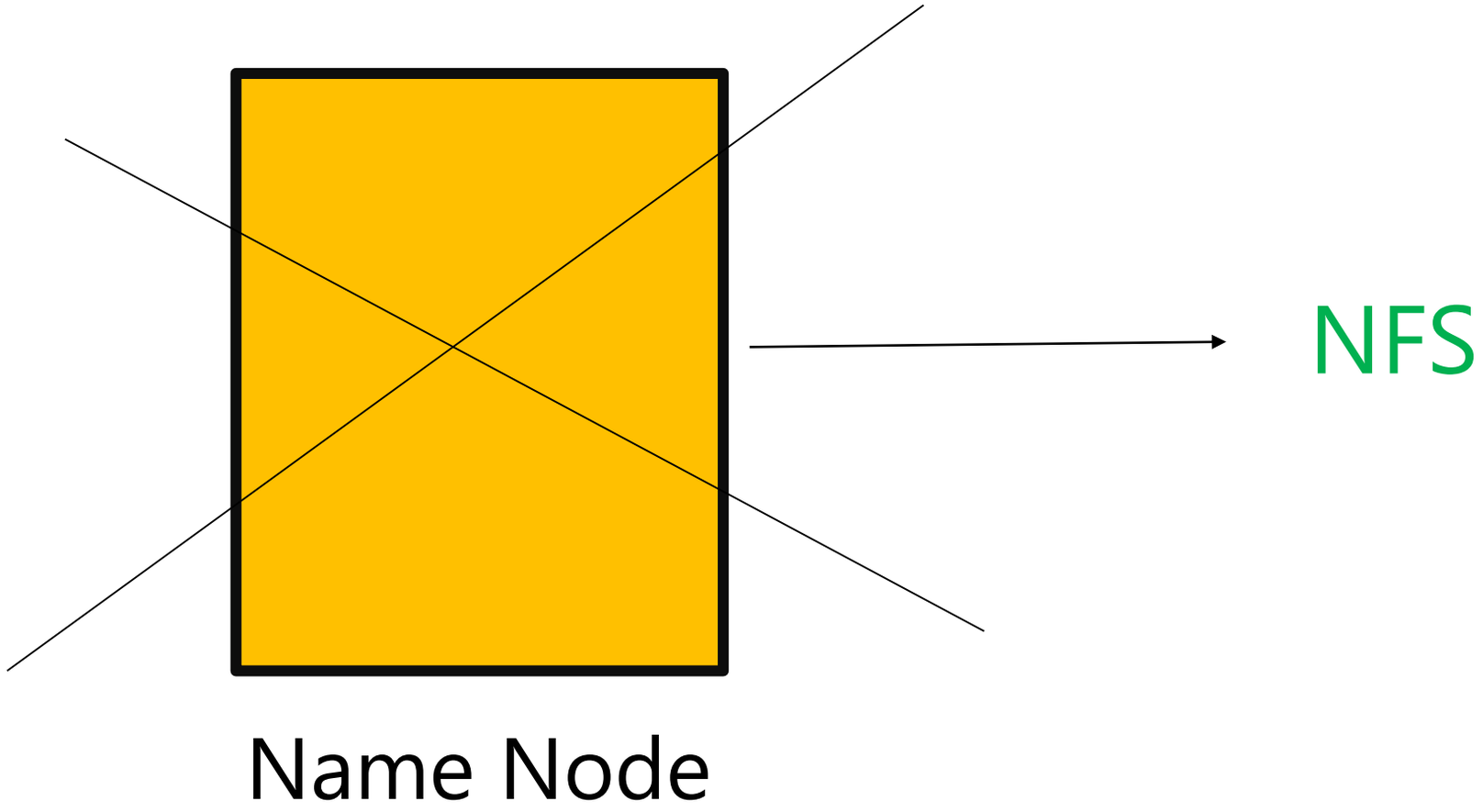


64 MB

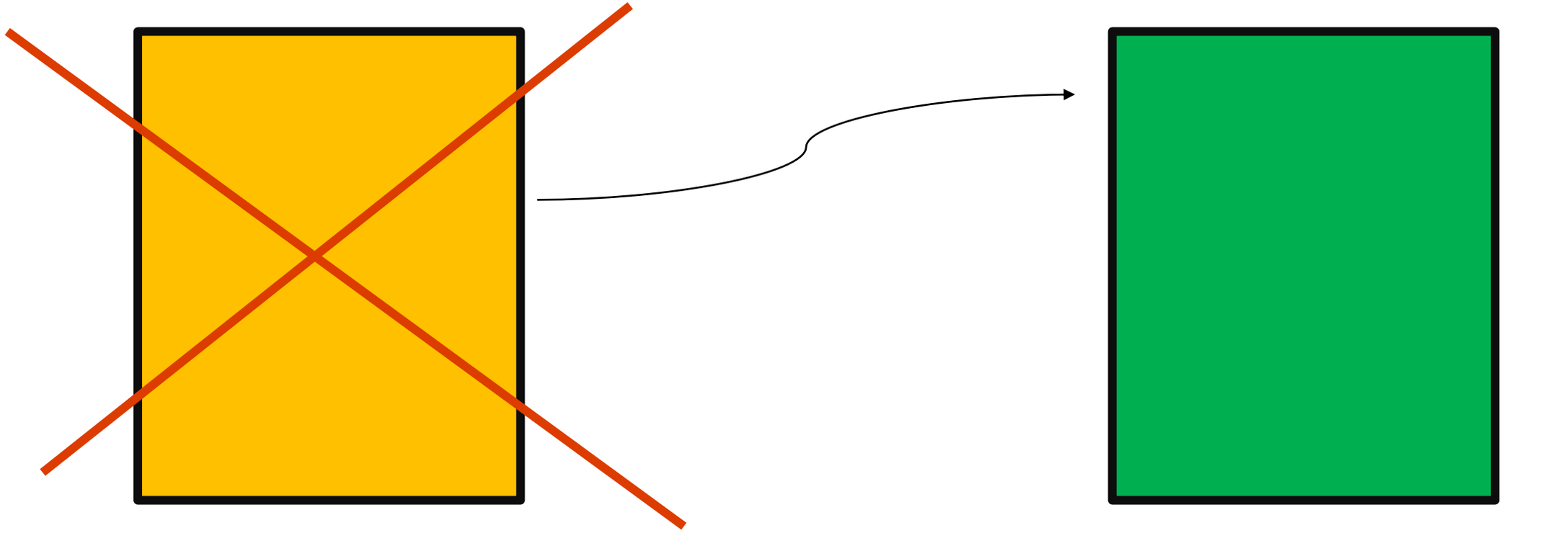
64 MB

22 MB

Name Node High Availability :



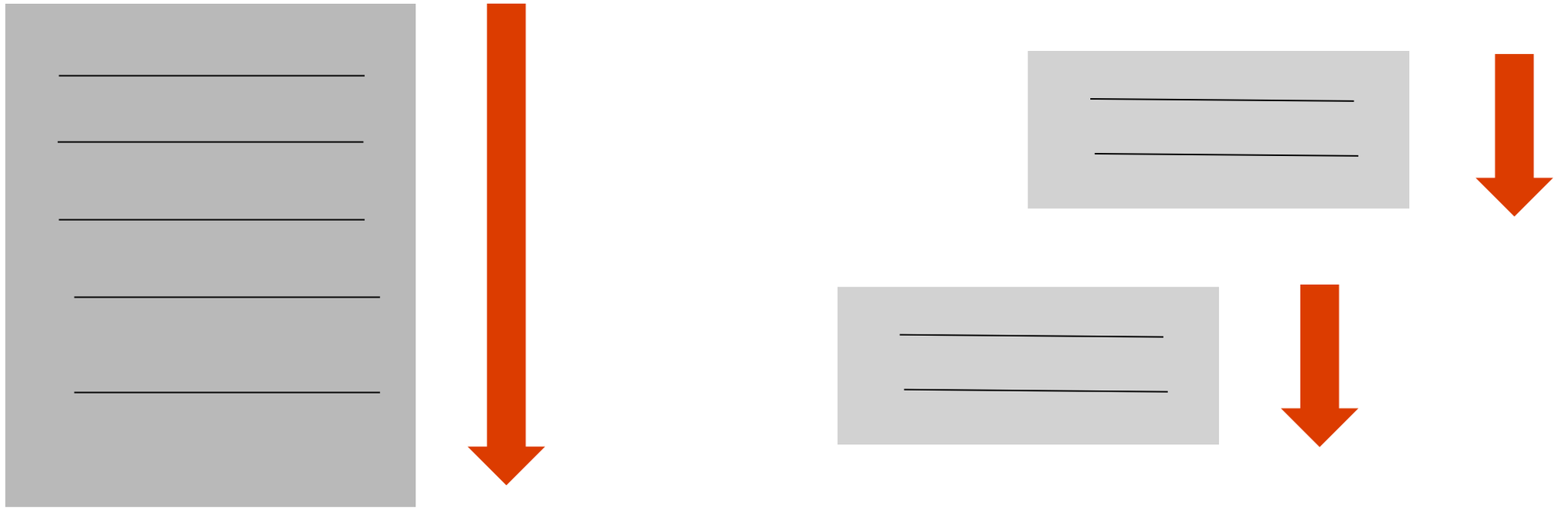
Active and Standby Name Node :



DEMO

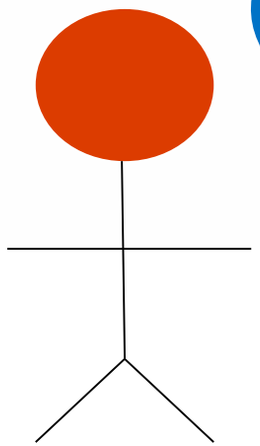
Inserting files in HDFS and performing various Operations on the files in HDFS

Map Reduce :



Map Reduce :


```
2012-01-01  London Clothes 25.99
2012-01-01  Miami  Music  12.15
2012-01-02  NYC  Toys  3.10
2012-01-02  Miami  Clothes 50.00
```



Calculate Total
Sales Per Store ?

Map Reduce :

2012-01-01 London Clothes 25.99
2012-01-01 Miami Music 12.15
2012-01-02 NYC Toys 3.10
2012-01-02 Miami Clothes 50.00

Location	Amount
London	25.99

Map Reduce :

2012-01-01 London Clothes 25.99
2012-01-01 Miami Music 12.15
2012-01-02 NYC Toys 3.10
2012-01-02 Miami Clothes 50.00

Location	Amount
London	25.99
Miami	12.15
NYC	3.10

Map Reduce :

2012-01-01 London Clothes 25.99
2012-01-01 Miami Music 12.15
2012-01-02 NYC Toys 3.10
2012-01-02 Miami Clothes 50.00

Continue

Location	Amount
London	25.99
Miami	62.15
NYC	3.10

Map Reduce :

Solution => HashTables

Location	Amount
London	25.99
Miami	62.15
NYC	3.10

Key : Value

Map Reduce :

Solution => HashTables

Location	Amount
London	25.99
Miami	62.15
NYC	3.10

Key : Value

Problems if Running on 1TB of Data ?

Map Reduce : Hash Tables

Problems if Running on 1TB of Data ?

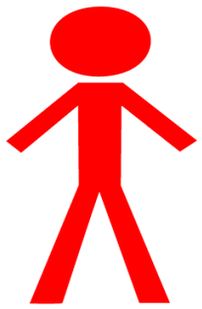
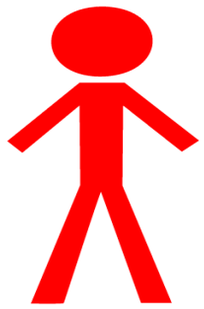
- Run Out of Memory
- Long Time

Location	Amount
London	25.99
Miami	62.15
NYC	3.10

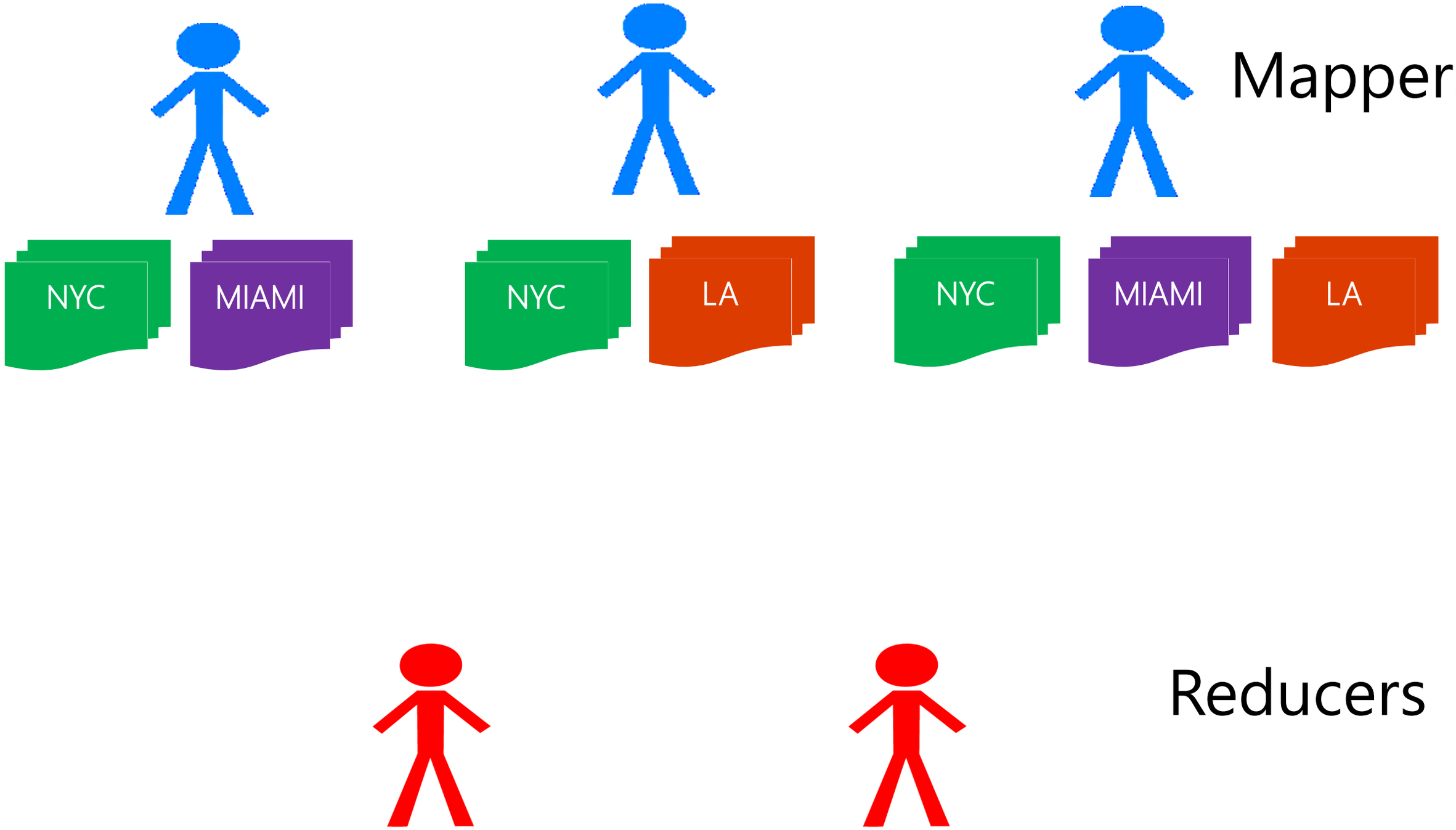
Key : Value

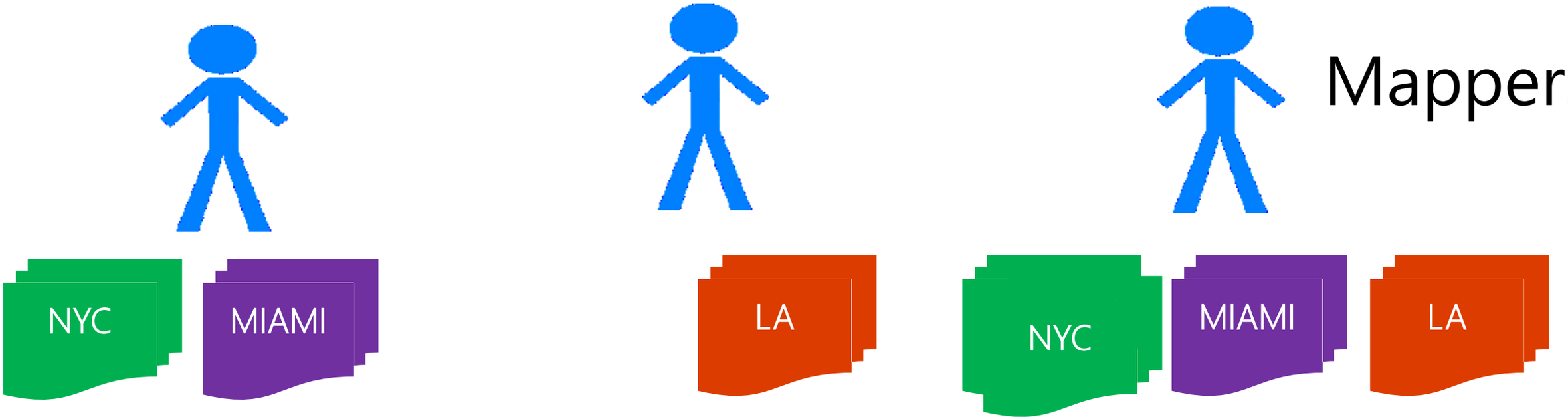


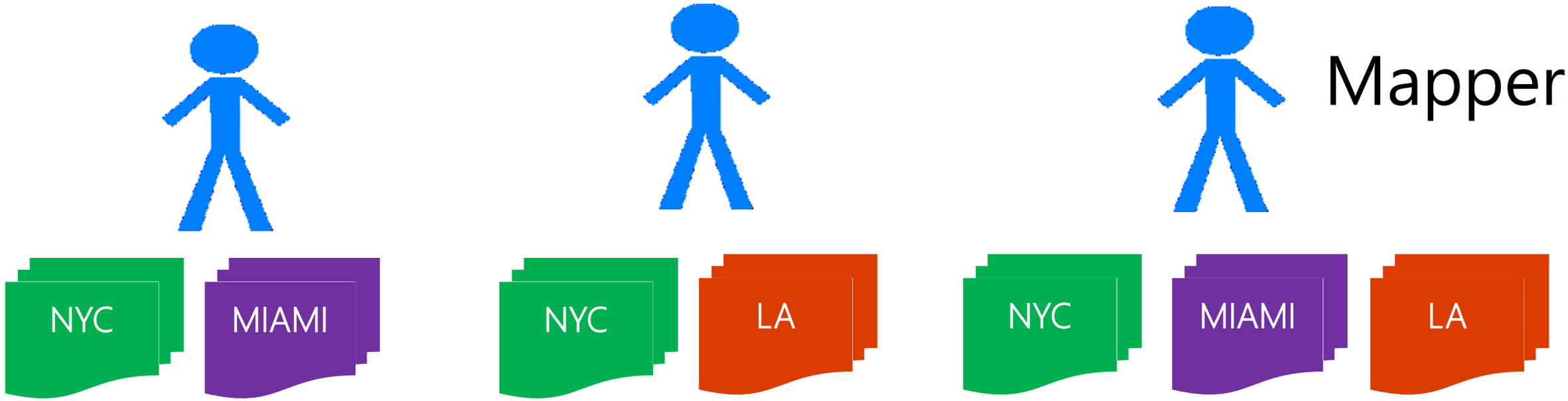
Mappers

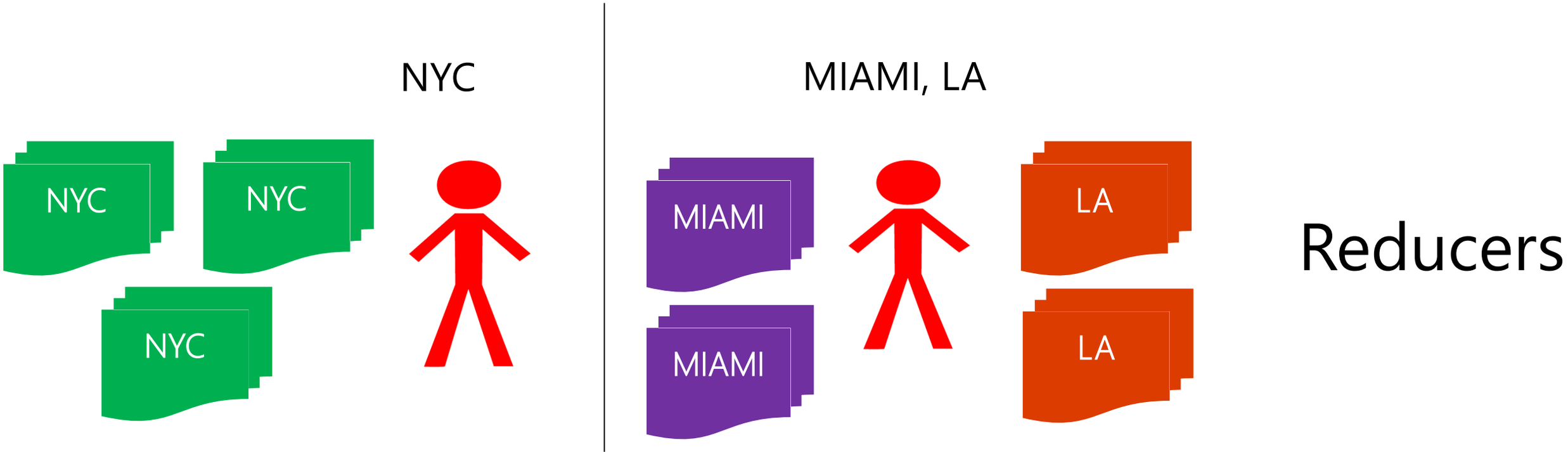



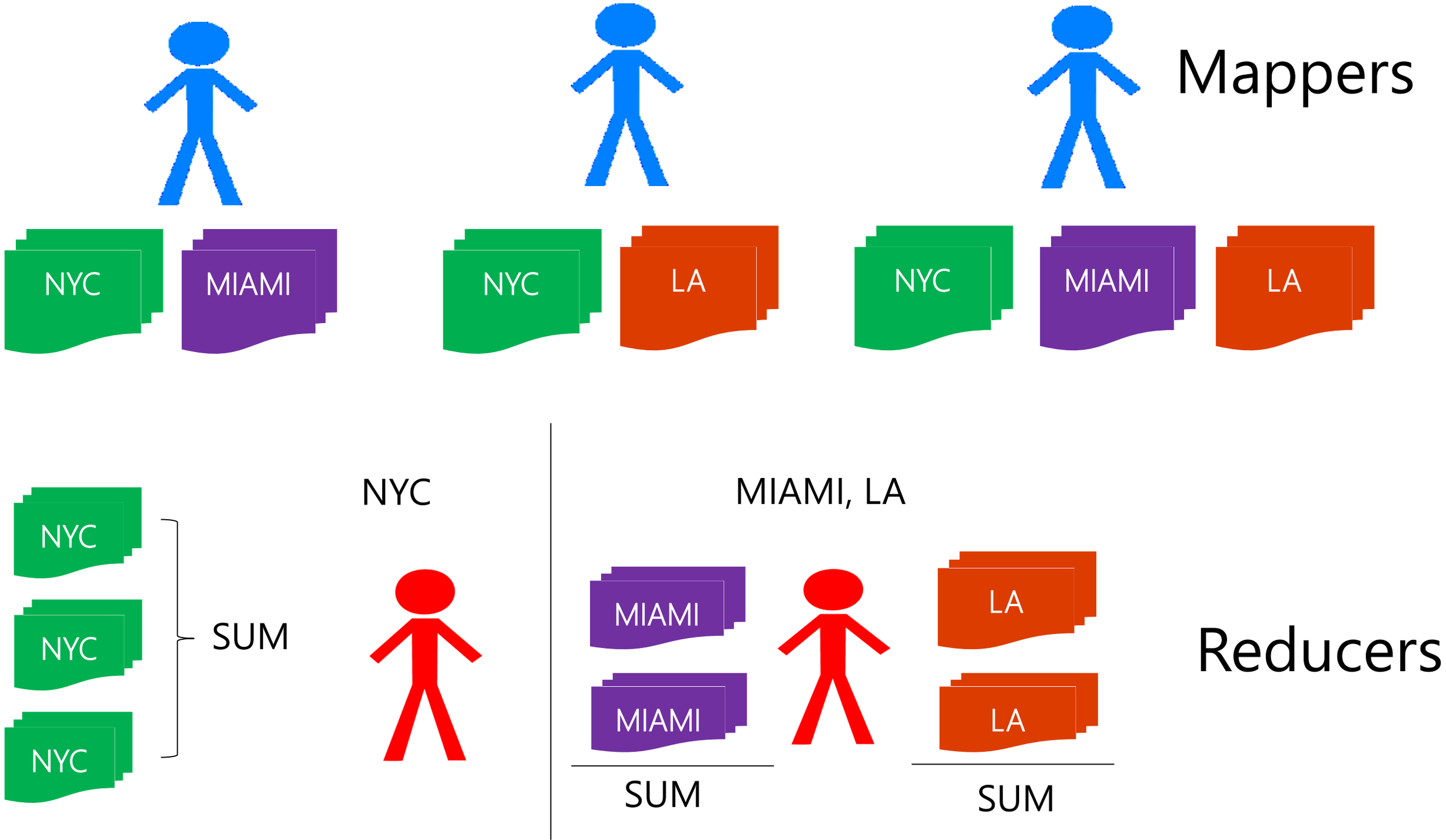
Reducers

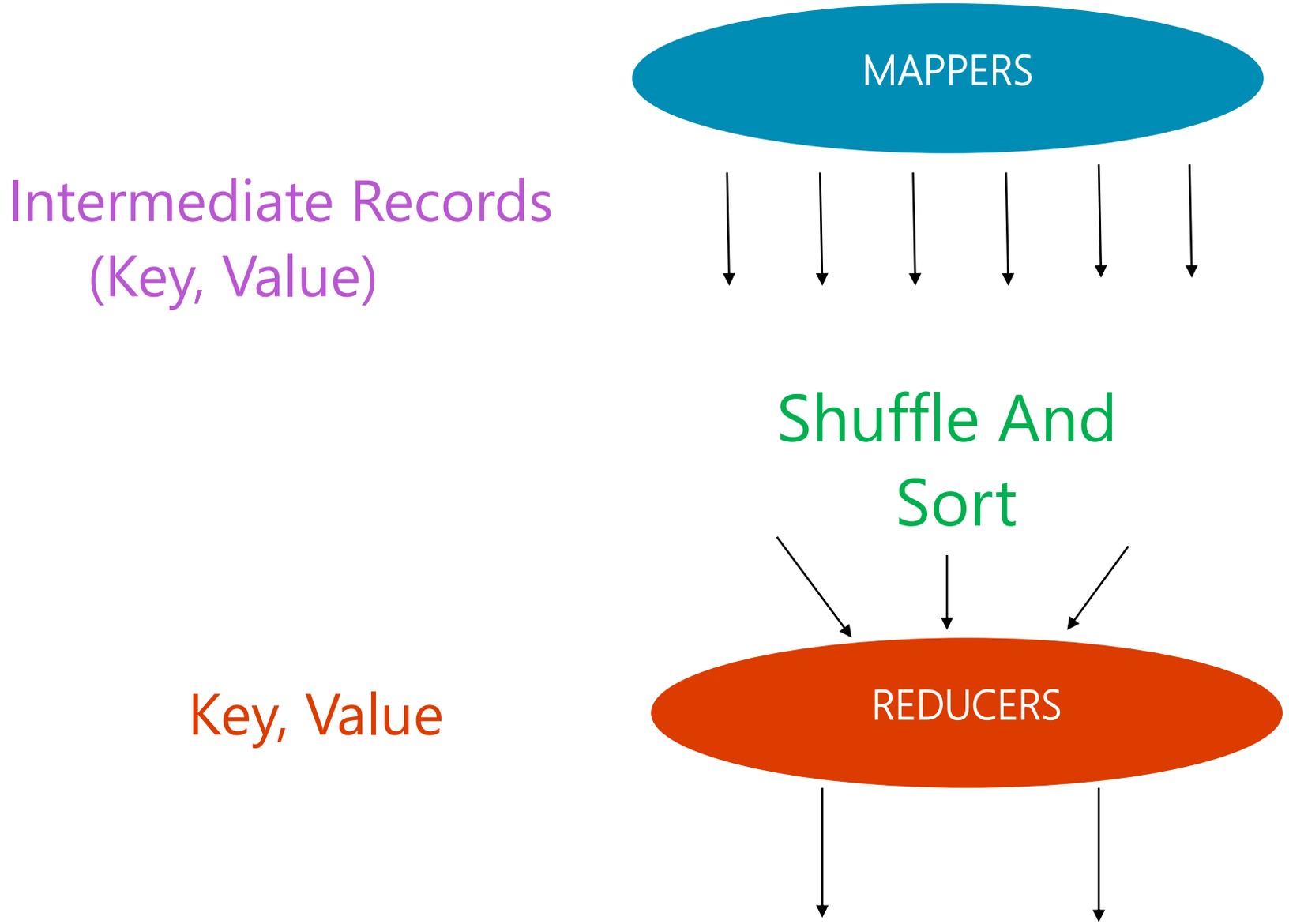

Mappers
NYC
MIAMI
NYC
LA
NYC
MIAMI
LA
Reducers


Mappers
NYC
MIAMI
LA
NYC
MIAMI
LA
NYC
MIAMI, LA
Reducers

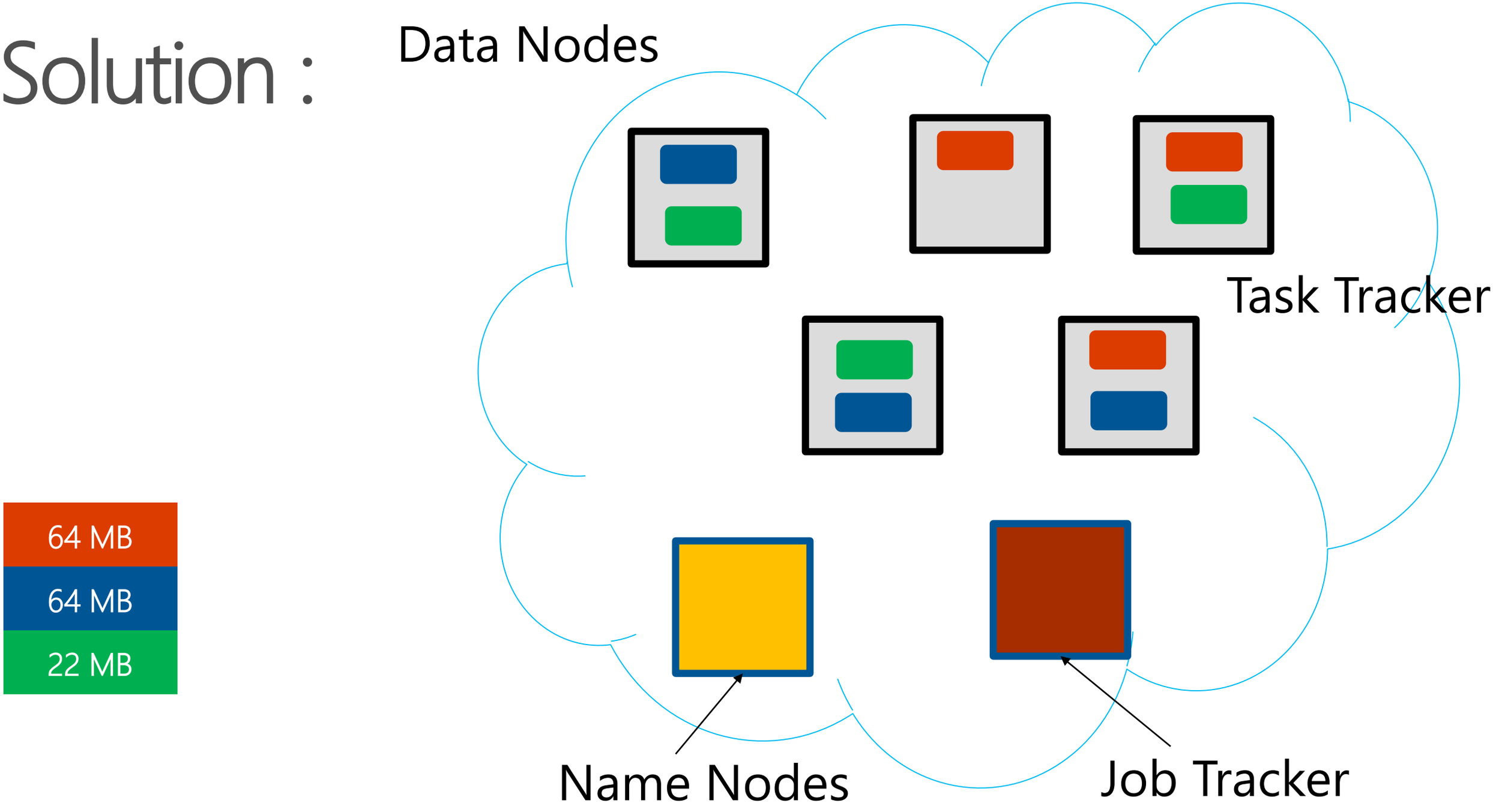

NYC
MIAMI
NYC
LA
NYC
MIAMI
LA
Mappers


NYC
NYC
NYC
NYC
MIAMI, LA
MIAMI
MIAMI
LA
LA
Reducers


Mappers
NYC
MIAMI
NYC
LA
NYC
MIAMI
LA
NYC
NYC
NYC
SUM
NYC
MIAMI, LA
MIAMI
MIAMI
LA
LA
SUM
SUM
Reducers

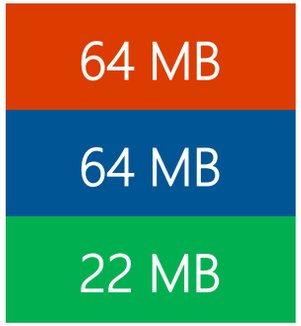

MAPPERS
Intermediate Records
(Key, Value)
Shuffle And
Sort
Key, Value
REDUCERS
RESULT

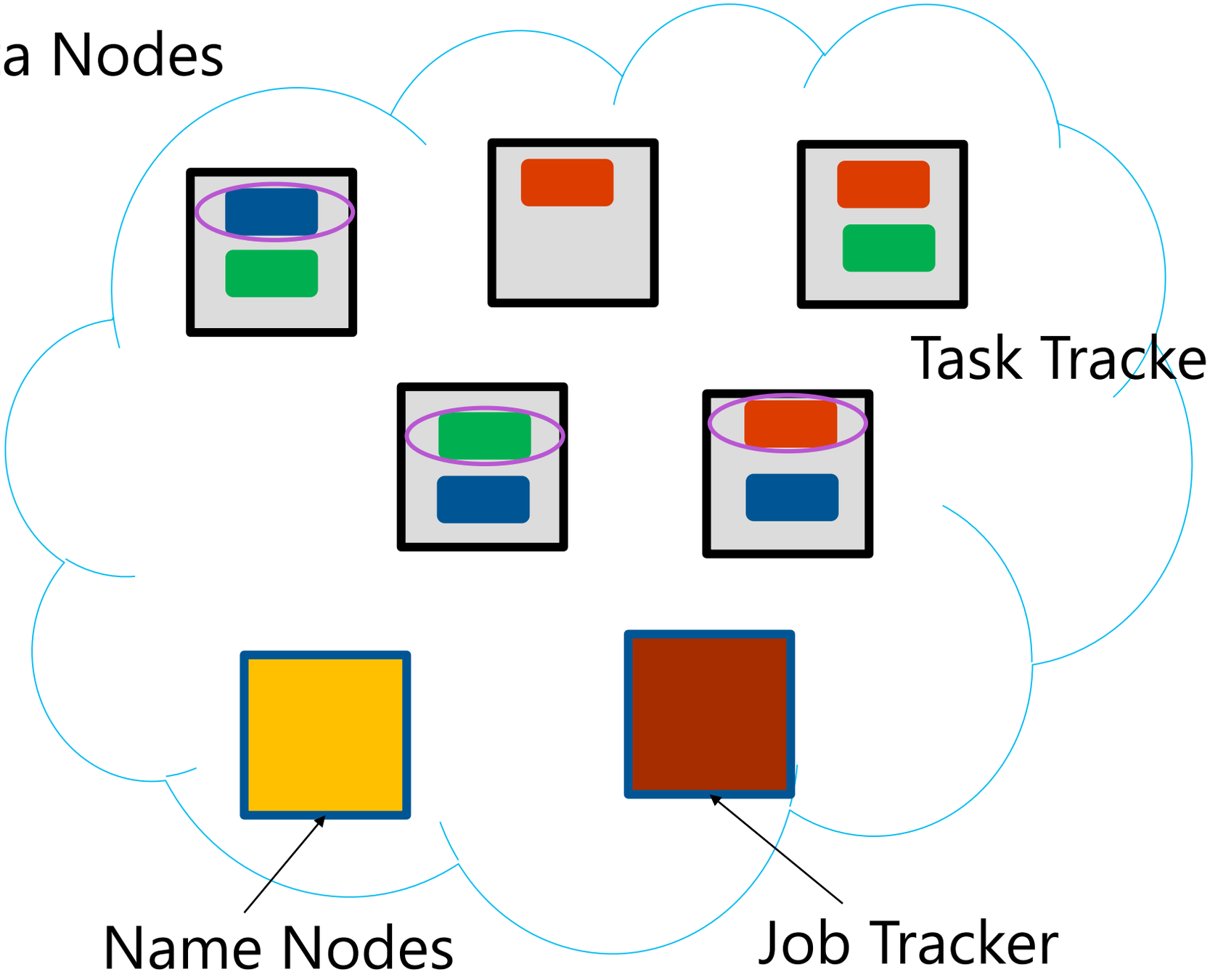
Solution :


Data Nodes
Task Tracker
64 MB
64 MB
22 MB
Name Nodes
Job Tracker

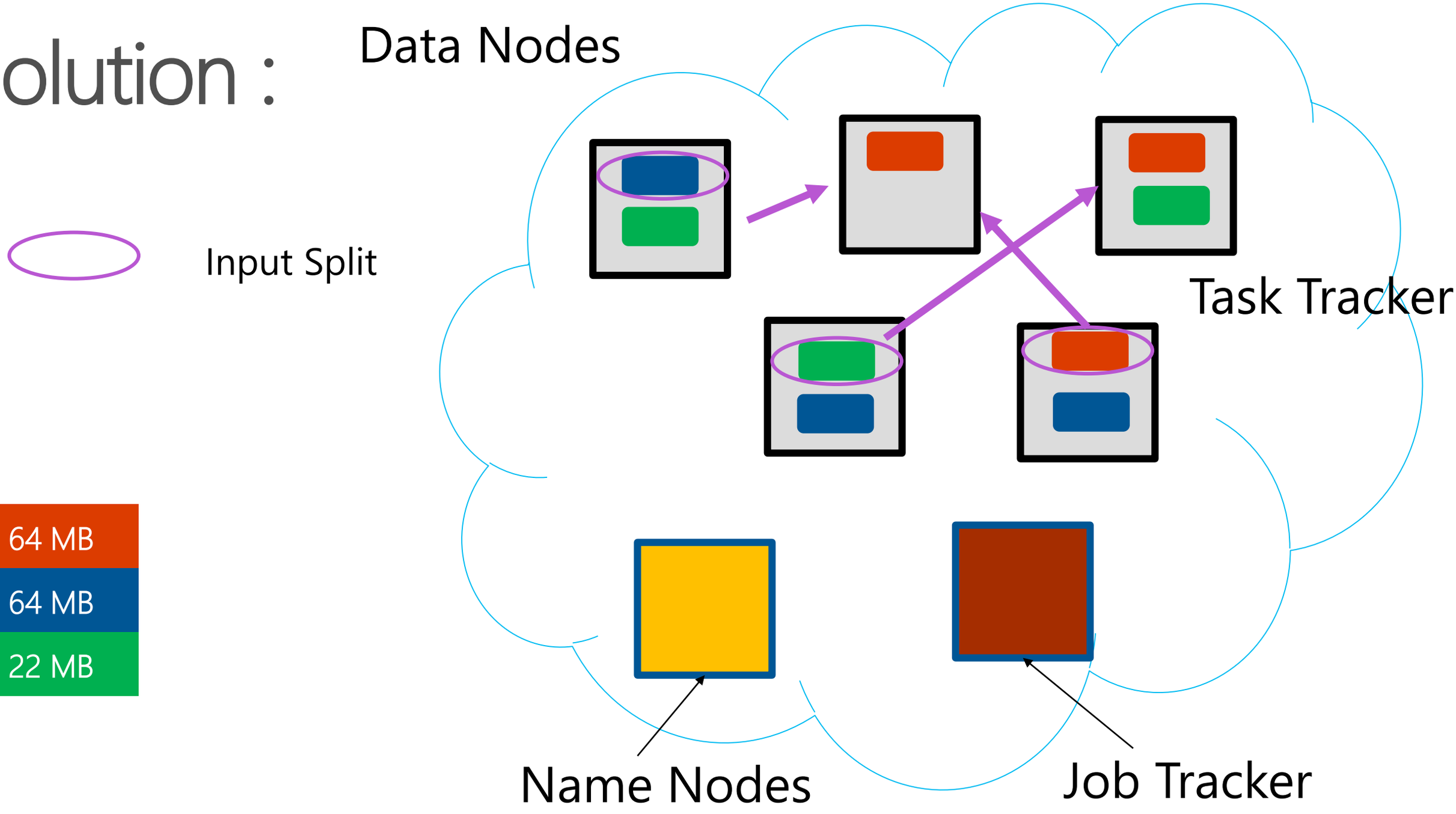
Solution :

Input Split





Solution :


Data Nodes
Input Split
64 MB
64 MB
22 MB
Task Tracker
Name Nodes
Job Tracker

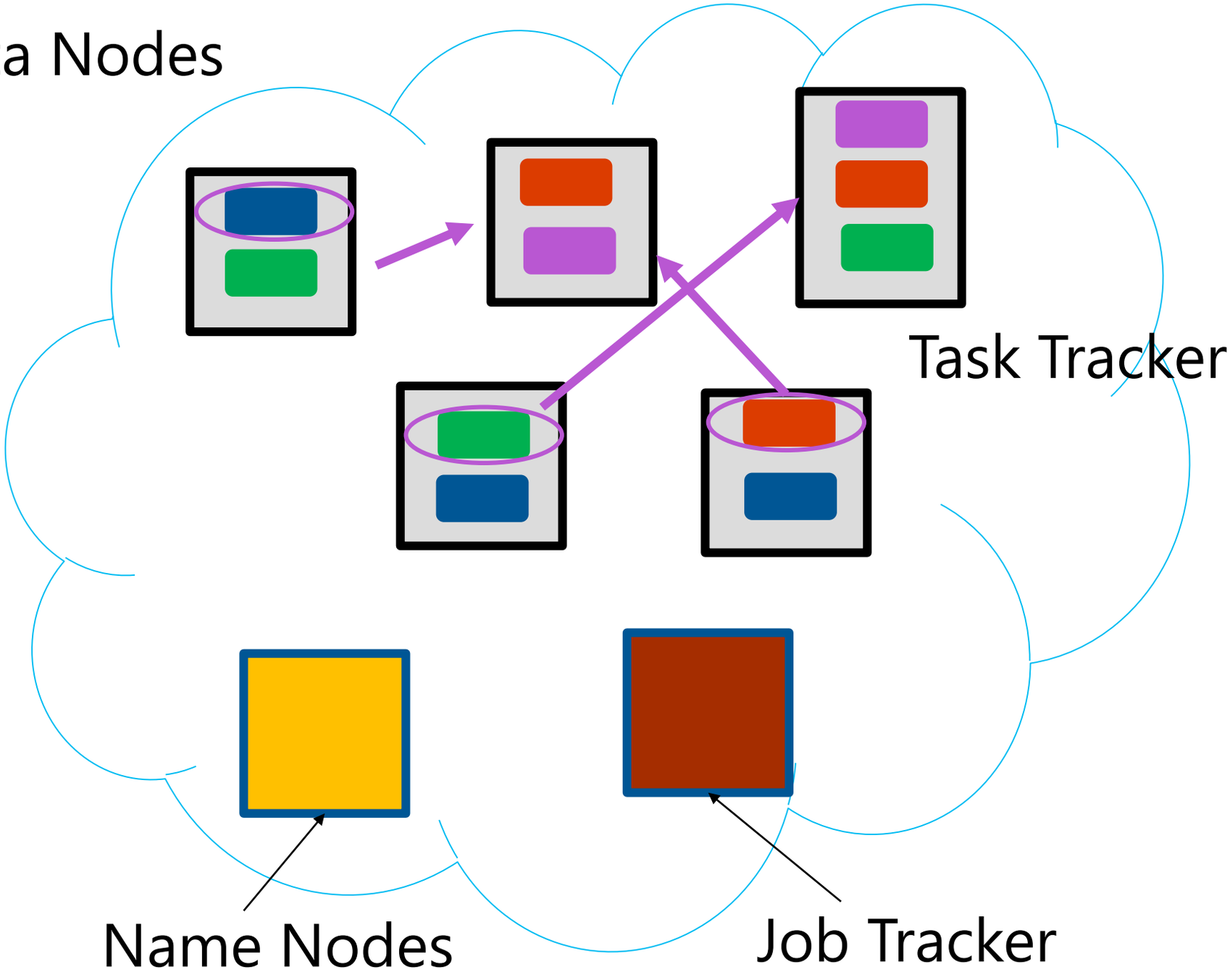
Solution :

Input Split

64 MB

64 MB

22 MB



Daemons of Map Reduce:

- A job is submitted to a **Job Tracker** which splits the work into Mappers and Reducers.
- Those mapper and reducers runs on the other data nodes.
- Running the actual Mar reduce task is handle by a daemon called **Task Tracker.**
- The Task Tracker Software will run on each of this nodes

Daemons of Map Reduce:

- As the task tracker runs on the same machine as the data node, the Hadoop framework will be able to help the map task work directly on piece of data that are stored on that machine.
- This will save a lot of network traffic.

DEMO

Running the Mapper and Reducer code in Hadoop.

Mapper Code :

```
def mapper():
    for line in sys.stdin:
        data = line.strip().split("\t")
        date, time, store, item, cost, payment = data
        print "{0}\t{1}".format(store, cost)
```

Mapper Code :

```
2012-01-01  12:01   San Jose    Music   12.99   Amex
2012-01-02  There was an error trying to connect to the database. Please try again.
```

```
def mapper():
    for line in sys.stdin:
        data = line.strip().split("\t")
        date, time, store, item, cost, payment = data
        print "{0}\t{1}".format(store, cost)
```

Mapper Code :

```
def mapper():
    for line in sys.stdin:
        data = line.strip().split("\t")

        if len(data) == 6:

            date, time, store, item, cost, payment = data

            print "{0}\t{1}".format(store, cost)
```

Reducer Code :

```
Miami   12.34
Miami   99.07
Miami   3.14
NYC     99.77
NYC     88.99
```

```
def reducer():

    salesTotal = 0
    oldKey = None

    for line in sys.stdin:
        data = line.strip().split("\t")

        if len(data) != 2:
            continue

        thisKey, thisSale = data

        if oldKey and oldKey != thisKey:
            print "{0}\t{1}".format(oldKey, salesTotal)

            salesTotal = 0

        oldKey = thisKey
        salesTotal += float(thisSale)
```

Reducer Code :

```
Miami   12.34
Miami   99.07
Miami   3.14
NYC     99.77
NYC     88.99
```

```
def reducer():

    salesTotal = 0
    oldKey = None

    for line in sys.stdin:
        data = line.strip().split("\t")

        if len(data) != 2:
            continue

        thisKey, thisSale = data

        if oldKey and oldKey != thisKey:
            print "{0}\t{1}".format(oldKey, salesTotal)

            salesTotal = 0

        oldKey = thisKey
        salesTotal += float(thisSale)

    if oldKey != None:
        print "{0}\t{1}".format(oldKey, salesTotal)
```