Machine Learning Using Big Data

Ashish Tripathi
JIIT, Noida
mail2ashish07@gmail.com

Big Data Analytics

Data Science and Analytics

Large-Scale Data Management

Introduction to Big Data

AMAZING FACTS

THE WORLD'S LARGEST 'BIG DATA' COMPANY GOOGLE

processes over **40,000 SEARCH QUERIES EVERY SECOND**, on average

That's over **3.5 BILLION** searches per day and **1.2 TRILLION** searches per year worldwide

LOS ANGELES POLICE DEPT

uses the big data model developed by **Professor George Moher** to **bring down crime rates** in the LA metropoiltan area

It allows the **LAPD** to identify and predict which areas will be crime hotspots in the future

So far it has **REDUCED...**

BURGLARIES by **33%**

21% **VIOLENT CRIMES** by

and **PROPERTY CRIMES** by **12%**

Source: Crayon Data

Google

stores an estimated

10 EXABYTES

of data.

Amazon hosts an estimated **1 exabyte** of data across more than **1,400,000 SERVERS**, making it the company with the most number of servers.

ONE EXABYTE

is a quintillion bytes, or roughly **1,600,000,000,000 books**, that's just **10,000 miles short of reaching the moon.**

Source: Adeptia

Zynga collects

25 TERABYTES

a day from Farmville to drive higher in-game purchases

Source: Crayon Data

Data is projected to grow into a

$53.4 billion

market by 2017

Source: BaselineMag.com

Every **query has to travel** on average

1,500 MILES

to a data center and back to return the answer to the user

A single **Google query** uses

1,000 COMPUTERS

in **0.2 seconds**

to retrieve an answer

Source: Internetlivestats.com

90% of the world's data was created in the last two years.

Source: Nasscom & Hansa Cequity report on Indian Analytics

3.2

ZETTABYTES

the **SIZE OF THE DIGITAL UNIVERSE IN 2014.**

40

ZETTABYTES

the estimated size of the digital universe in 2020

(1 zettabyte = 1,024 exabytes.)

INDIA SPECIFIC INFO

LESS THAN 20%

of **INDIAN ORGANISATIONS** consider Big Data as **"HYPE"** or the latest technology buzzword.

Source - Ernst and Young

91% of Indian businesses agree that big data will lead to better decision making.

AWARENESS

about big data is pretty high among Indian enterprises, with **70%** saying that it will be a key factor in determining winners and losers in their industry.

Source: EMC

50% of large organisations plan to use data from social networks for sentiment analysis and customer tracking.

Source: Ernst & Young

75% of organisations are confident of **DRIVING NEW REVENUE STREAMS USING BIG DATA**, however, only **35%** of organizations **PLAN TO INVEST** in building Big Data capabilities

Source: EMC

ANALYTICS MARKET in India is expected to

GROW 22.8%

from 2014 to 2019

LOYAL CUSTOMERS ARE WORTH UP TO

10 TIMES

THEIR FIRST PURCHASE

Top-performing organisations show a

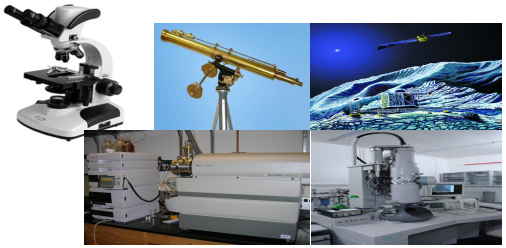
5X HIGHER USAGE

of analytics as opposed to low performers

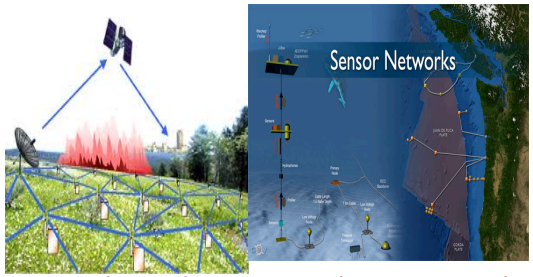
Source: Nasscom & Hansa Cequity report on Indian Analytics

Who's Generating Big Data

Social media and networks
(all of us are generating data)



Scientific instruments
(collecting all sorts of data)



Sensor technology and networks
(measuring all kinds of data)

Mobile devices
(tracking all objects all the time)

Source:http://www.slideshare.net/kmstechnology/big-data-overview-2013-2014
http://www.slideshare.net/sajjanvsl/final-presentation-45456729

The Model Has Changed…

- **The Model of Generating/Consuming Data has Changed**

Old Model: Few companies are generating data, all others are consuming data

New Model: all of us are generating data, and all of us are consuming data

Definition

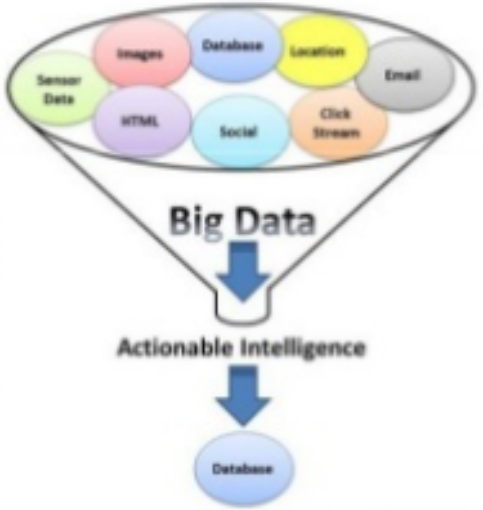
- Introduction to Big data
- 3 V's of Big Data
- Volume
- Velocity
- Variety

Introduction to Big Data

- Big Data Definition
- No single standard definition…

Ref: goo.gl/iWZhjJ

Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.



http://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/#379879e621a9

Popular Tools For Processing Big Data

- Hadoop/ MapReduce
- Pig
- Hive
- Mahout
- Spark

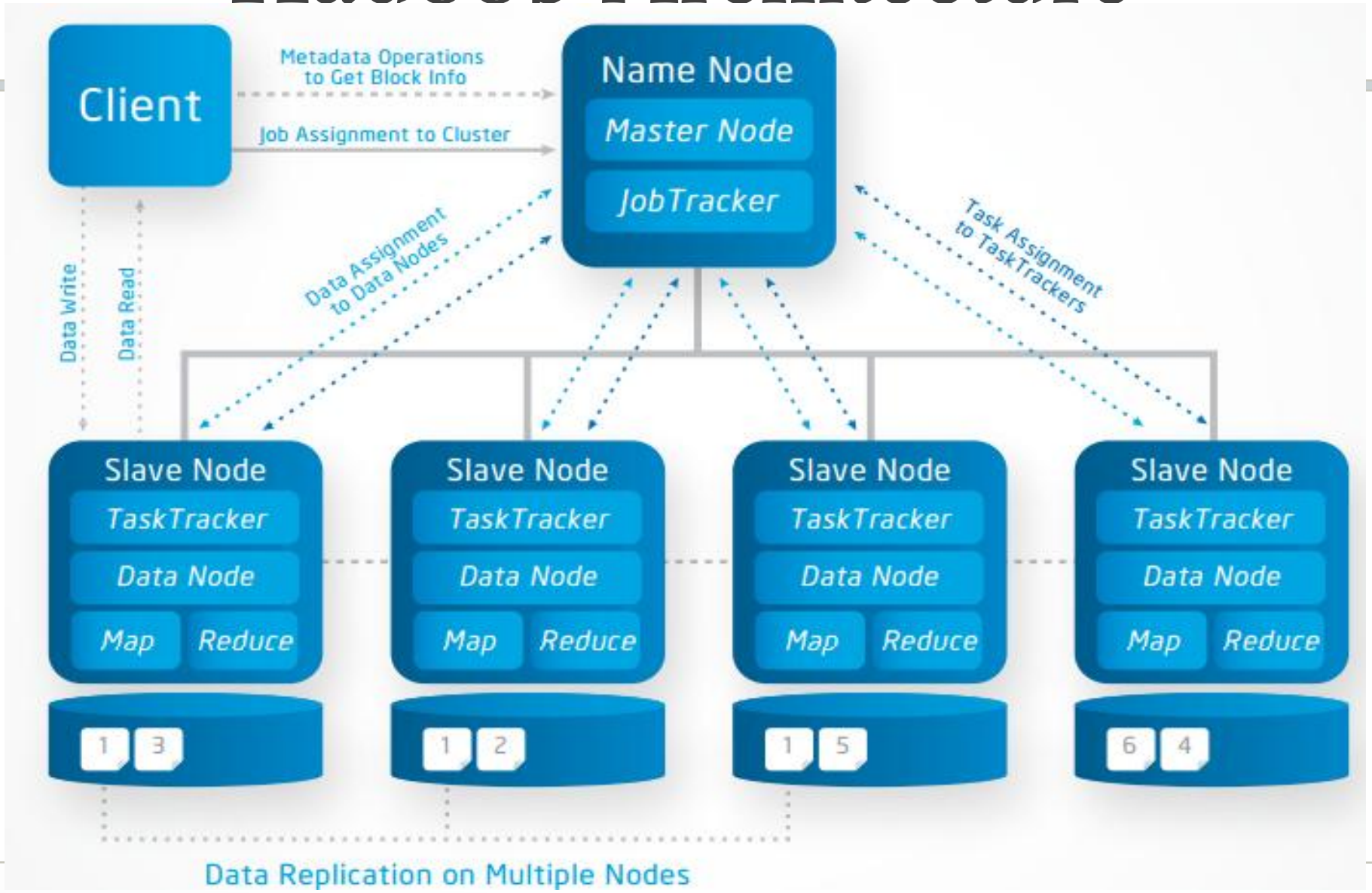
Hadoop

Apache Hadoop is an open source framework or software library which provide platform for running for the distributed computation on the cluster built of commodity hardware.

- Hadoop is designed to distribute large amounts of work across a set of machines called cluster.

MapReduce is a programming model for processing the large datasets which runs on the top of hadoop.

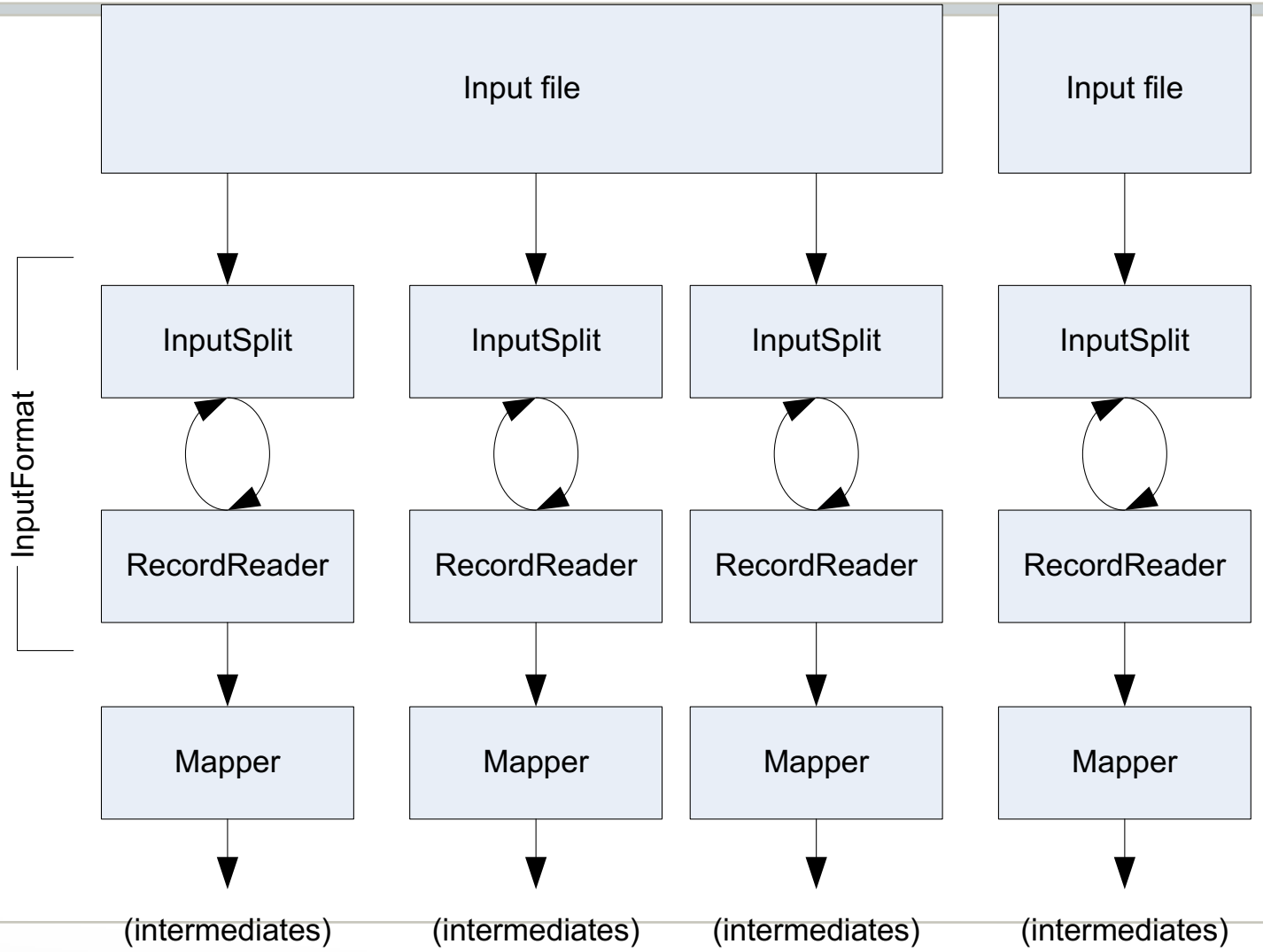
Hadoop Architecture



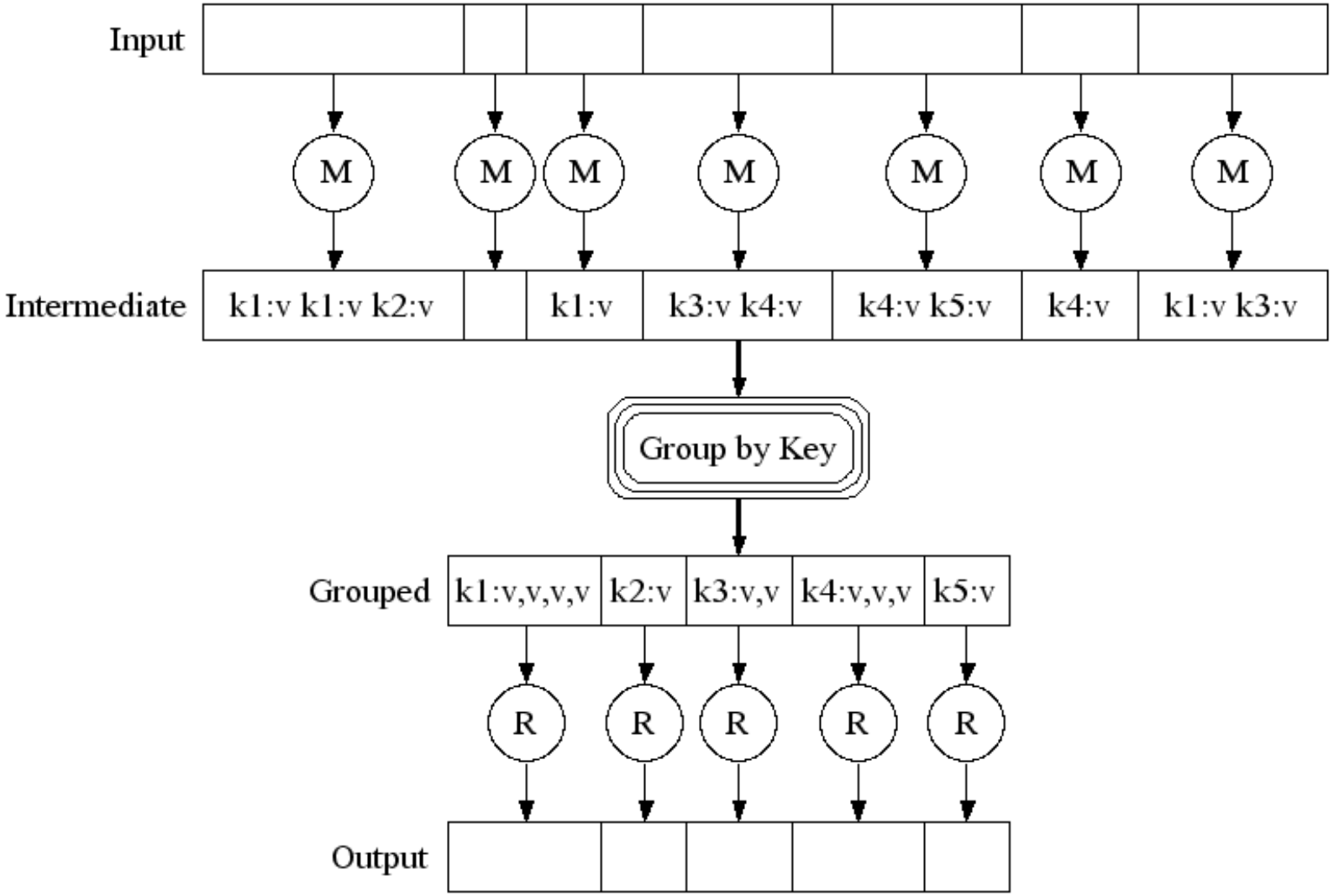
Why So popular

- Uses Commodity hardware
- Reliable
- Scalable

Getting Data To The Mapper



The MapReduce Framework (pioneered by Google)



MapReduce in Hadoop (1)

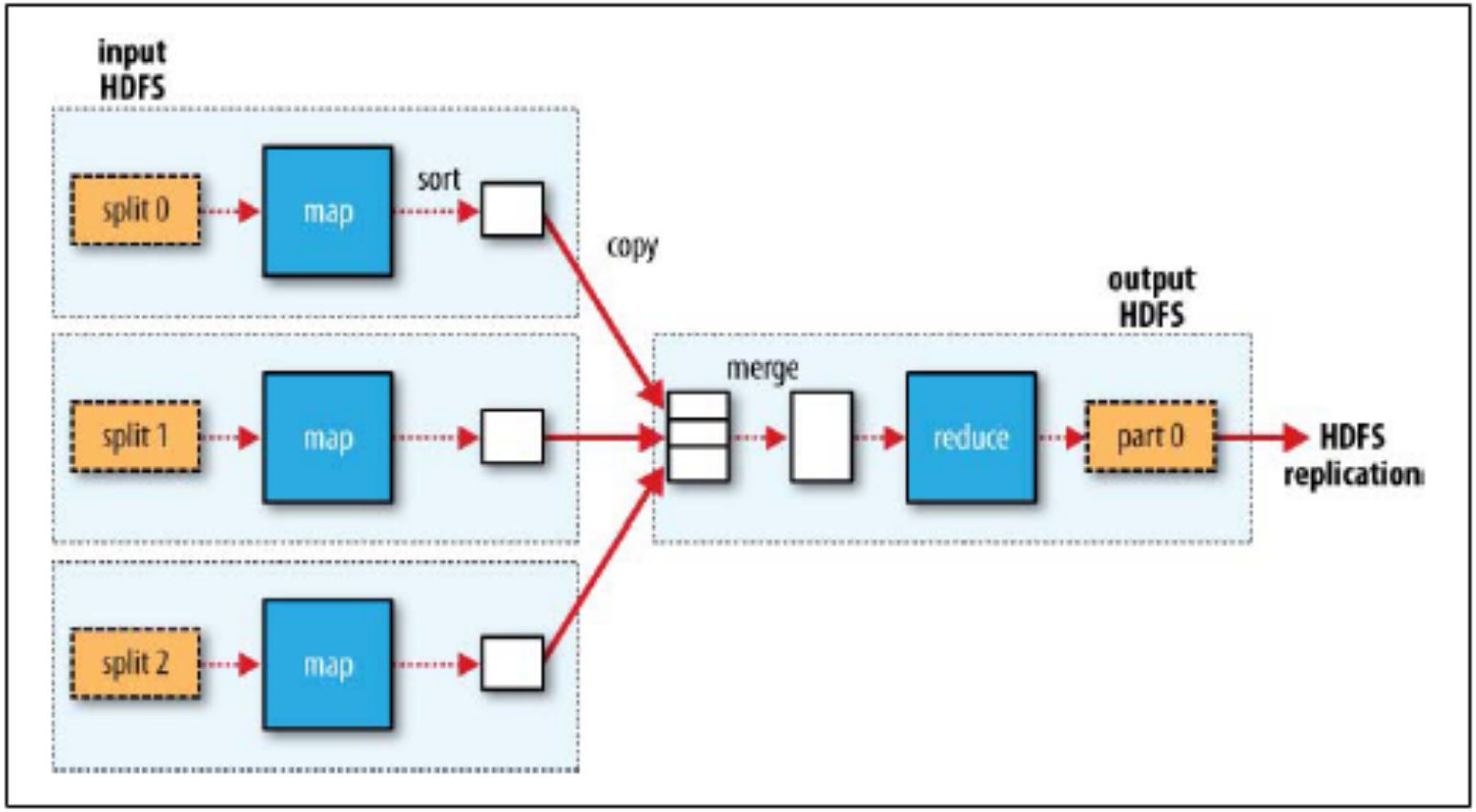


Figure 2-2. MapReduce data flow with a single reduce task

MapReduce in Hadoop (2)

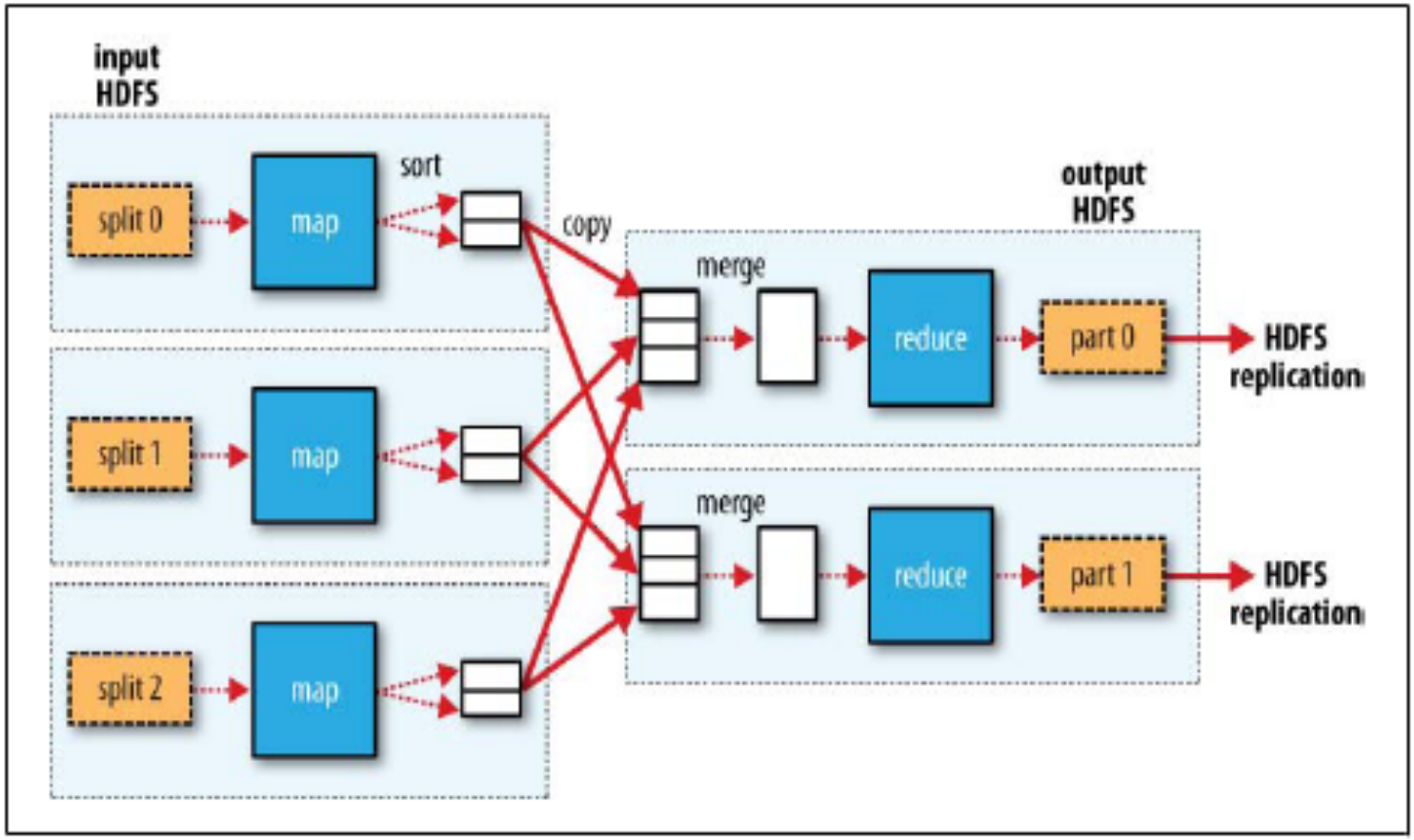


Figure 2-3. MapReduce data flow with multiple reduce tasks

MapReduce in Hadoop (3)

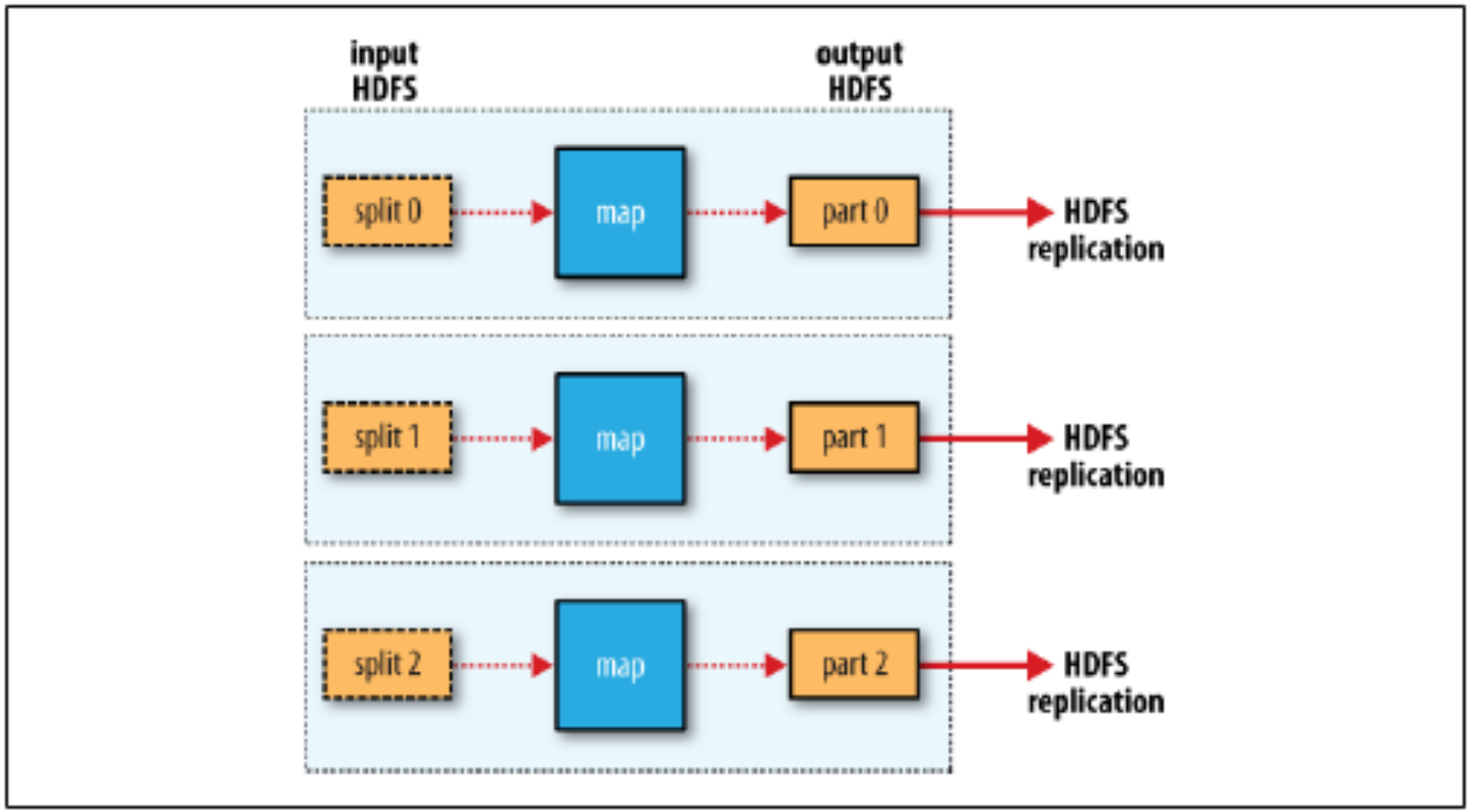
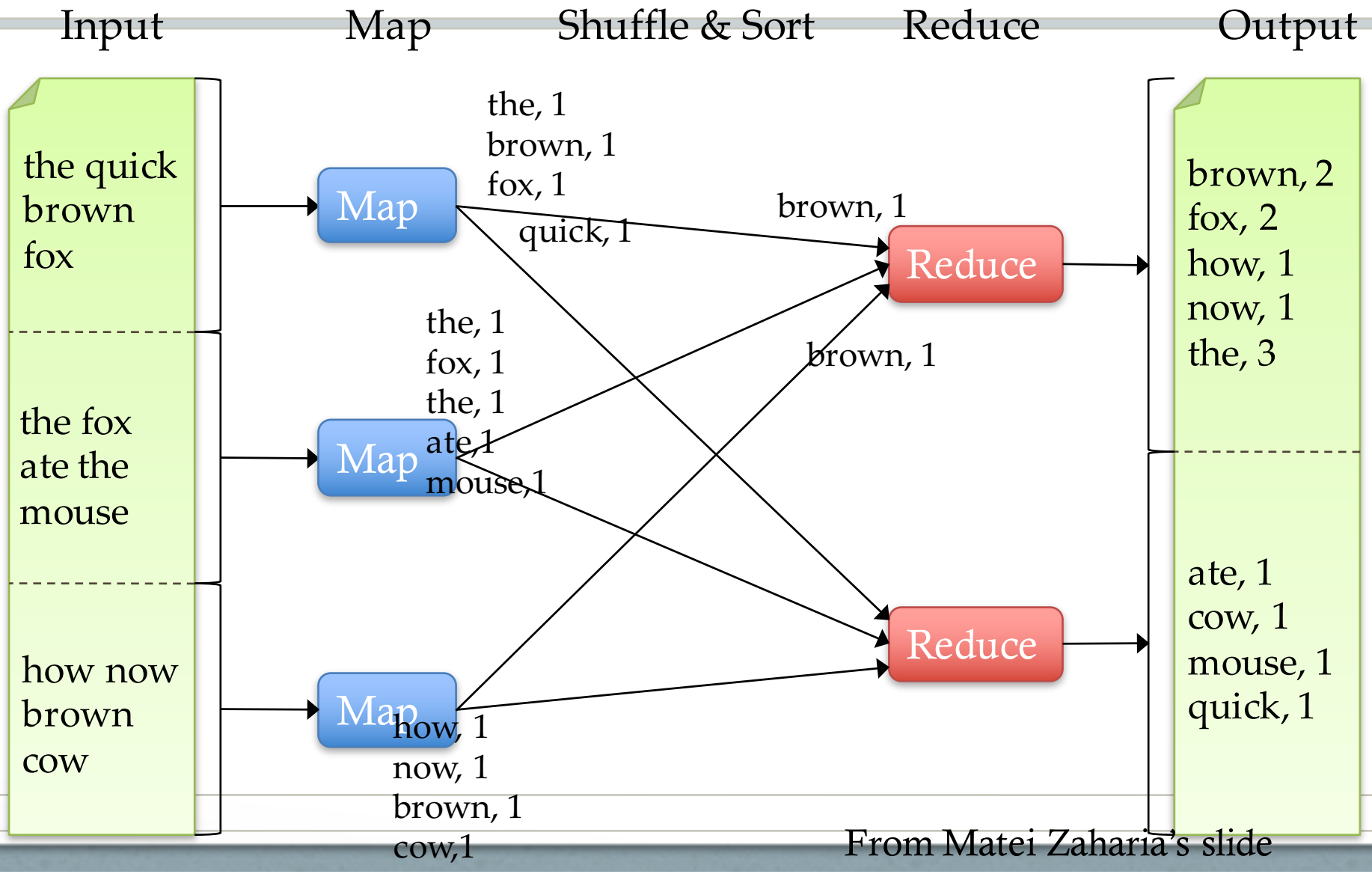


Figure 2-4. MapReduce data flow with no reduce tasks

Word Count Execution



From Matei Zaharia's slide

MapReduce WordCount.java

(Mapper Code)

```
public static class Wordcount extends MapReduceBase implements Mapper <LongWritable,
    Text, Text, IntWritable>{

  public void map(Long Writable key, Text value, OutputCollector<Text
    ,IntWritable>output ,Reporter   r) throws IOException{

    String s =value.toString());

    for(String word:s.split( " " )){                    //regular expression

            if(word.lenghth.>0)

            {       output.collect(new Text(word), new IntWritable(1))   //collect is a method of
    output collector interface

    }

  }
```

Reduce code in WordCount.java Reducer code

```
public class WordReducer  extends MapReduceBase Impliments
      Reducer<Text,IntWritable,Text,IntWritable> {

    public void reduce(Text key, Iterable<IntWritable> values,

                        OutputCollector<Text,Intwritable> output Reporter  ) throws IOException{

    int count =0;

   while(value.hasNext())   // this method checks that whether there is any value

{

       IntWritable i=value.Next();     //  this method get that value

       count +=i.get();                          // get method converts it to int(object type to primitive
      type.

                }
```

Reduce code in WordCount.java

```
Public class WordCounter {

Public static void main(String [] args) throws IOEXception, InterruptedException ,
      ClassNotFoundException{

  Job job =new Job();

  job.setJobName (“wordcounter ”);

  job.setJarByClass(WordCounter.class”);

 job.setMapperClass(Wordcount.class)

job.setReduceClass(Wordreduce.class)

Job.setOutputKeyClass(Text.class);

Job.setOutputValueClass(IntWritable.class);

FileInputFormat.addInputPath (Job, new Path(“/sample/word.txt”));
```

Steps to run job

1- Open mozila and write 50070 on address bar.

2- upload file to HDFS using command

hadoop fs -copyFromLocal input.txt /

3- run jar using command

hadoop jar wordcount.jar /input.txt /out/