

# Big Data Processing

## MapReduce

# Before MapReduce...

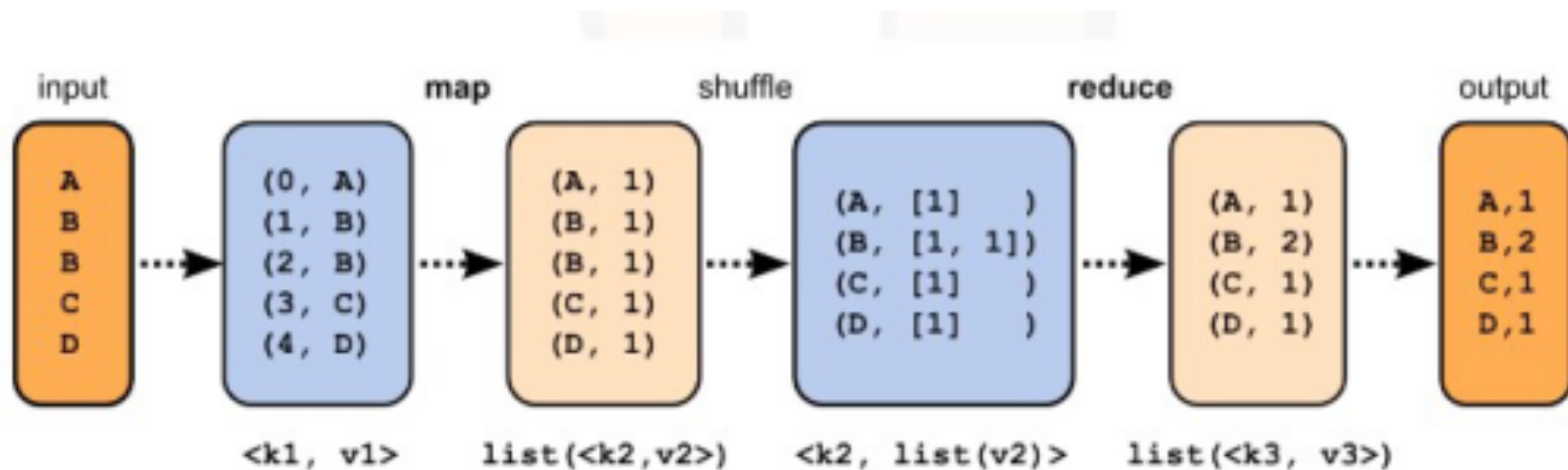
## Large scale data processing was difficult!

- Managing hundreds or thousands of processors
- Managing parallelization and distribution
- I/O Scheduling
- Status and monitoring
- Fault/crash tolerance

## MapReduce provides all of these, easily!

# How Map and Reduce Work Together

- Map returns information
- Reduces accepts information
- Reduce applies a user defined function to reduce the amount of data



# Example MapReduce: WordCount

**\$cd /guest1**



**\$wget https://github.com/bobbylovemovie/trainbigdata/raw/master/HDFS/  
wordcount.jar**

**\$hadoop jar wordcount.jar org.myorg.WordCount /user/cloudera/input/\*  
/user/cloudera/output/wordcount**

```
[cloudera@quickstart guest1]$ hadoop jar wordcount.jar org.myorg.WordCount /user/cloudera/input/* /user/cloudera/output/wordcount
16/10/12 23:46:17 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
16/10/12 23:46:17 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
16/10/12 23:46:18 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your applica
tion with ToolRunner to remedy this.
16/10/12 23:46:18 INFO mapred.FileInputFormat: Total input paths to process : 1
16/10/12 23:46:18 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:862)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:600)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:789)
16/10/12 23:46:18 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:862)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:600)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:789)
16/10/12 23:46:18 INFO mapreduce.JobSubmitter: number of splits:2
16/10/12 23:46:18 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1476334425555_0001
16/10/12 23:46:19 INFO impl.YarnClientImpl: Submitted application application_1476334425555_0001
16/10/12 23:46:19 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1476334425555_0001/
16/10/12 23:46:19 INFO mapreduce.Job: Running job: job_1476334425555_0001
16/10/12 23:46:32 INFO mapreduce.Job: Job job_1476334425555_0001 running in uber mode : false
16/10/12 23:46:32 INFO mapreduce.Job:  map 0% reduce 0%
16/10/12 23:46:54 INFO mapreduce.Job:  map 100% reduce 0%
16/10/12 23:47:06 INFO mapreduce.Job:  map 100% reduce 100%
16/10/12 23:47:07 INFO mapreduce.Job: Job job_1476334425555_0001 completed successfully
16/10/12 23:47:07 INFO mapreduce.Job: Counters: 49
File System Counters
```

# Reviewing MapReduce Job in Hue

**Job Browser** Jobs Workflows Schedules Bundles SLAs

user:cloudera ☐ Succeeded ☐ Running ☐ Failed in the last 7 days  

<input type="checkbox"/>	Id	Name	User	Type	Status	Progress	Group	Started	Duration
<input type="checkbox"/>	application_1504064742599_0001	wordcount	cloudera	MAPREDUCE	SUCCEEDED	100	root.cloudera	August 30, 2017 2:05 AM	1m, 0s

**Job Browser** Jobs Workflows Schedules Bundles SLAs

job\_1504064742599\_0001

ID  
job\_1504064742...

TYPE  
MAPREDUCE


STATUS  
SUCCEEDED

USER  
cloudera

PROGRESS  
100%


MAP  
100% 2 / 2

REDUCE  
100% 1 / 1

Logs **Tasks** Metadata Counters 

Filter by name ☐ Succeeded ☐ Running ☐ Failed ☐ Map ☐ Reduce




Type	Id	Elapsed Time	Progress	State	Start Time	Successful
MAP	task_1504064742599_0001_m_000000	28252	1	SUCCEEDED	1504083966966	attempt_1
MAP	task_1504064742599_0001_m_000001	27784	1	SUCCEEDED	1504083967448	attempt_1
REDUCE	task_1504064742599_0001_r_000000	11368	1	SUCCEEDED	1504083997798	attempt_1



# Reviewing MapReduce Output Result

## File Browser

[🏠 Home](#)[/ user / cloudera / output](#)[▼ History](#)

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	 <a href="#">↑</a>		cloudera	cloudera	drwxr-xr-x	August 30, 2017 02:06 AM
<input type="checkbox"/>	 <a href="#">.</a>		cloudera	cloudera	drwxr-xr-x	August 30, 2017 02:06 AM
<input type="checkbox"/>	 <a href="#">wordcount</a>		cloudera	cloudera	drwxr-xr-x	August 30, 2017 02:06 AM

Show  of 1 items

Page  of 1

# Reviewing MapReduce Output Result

## File Browser

View as  
binary

Home

Page 1 to 1 of 1



Edit file

Download

View file  
location

Refresh

Last modified  
08/30/2017  
9:06 AM

User  
cloudera

Group  
cloudera

Size  
44 B

/ user / cloudera / output / wordcount / **part-00000**

a	205807
e	315232
i	174282
o	192879
u	65433