

Big Data Processing

MapReduce

Before MapReduce...

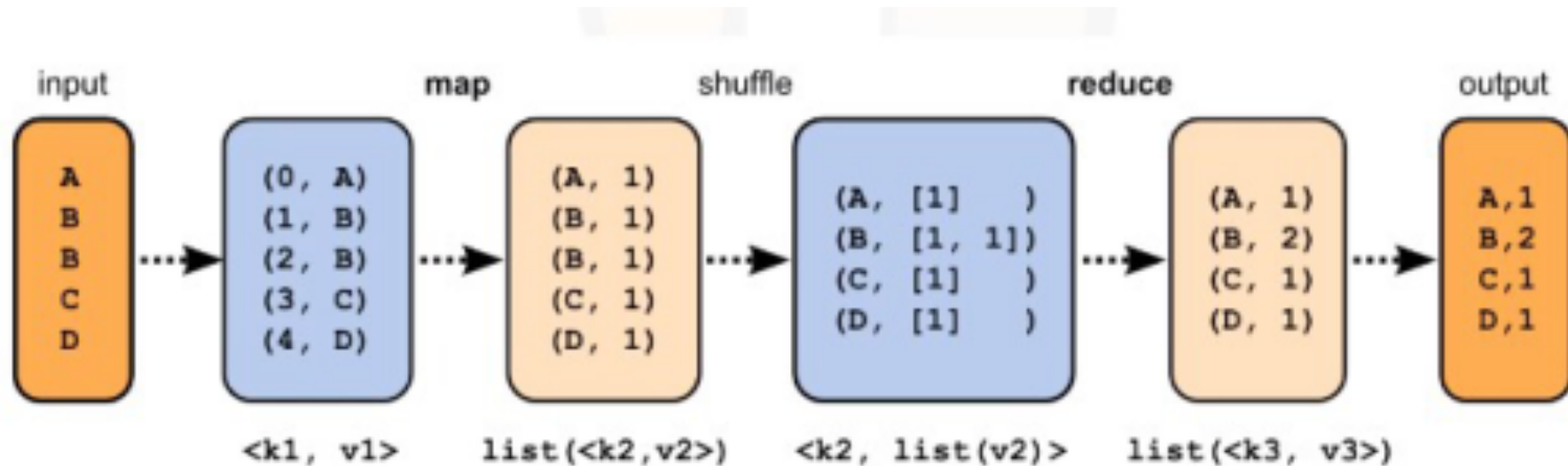
Large scale data processing was difficult!

- 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

HUE Home Query Editors ▾ Data Browsers ▾ Workflows ▾ Search Security ▾

Job Browser

Username cloudera

Text Search for text

Succeeded Running Failed Killed

Logs	ID	Name	Application Type	Status	User	Maps	Reduces	Queue	Priority	Duration	Submitted
	1476334425555_0001	wordcount	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	46s	10/12/16 23:46:19

HUE Home Query Editors ▾ Data Browsers ▾ Workflows ▾ Search Security ▾

Job Browser

JOB ID

1476334425555_0001

TYPE

MR2

USER

cloudera

STATUS

SUCCEEDED

wordcount

Attempts

Tasks

Metadata

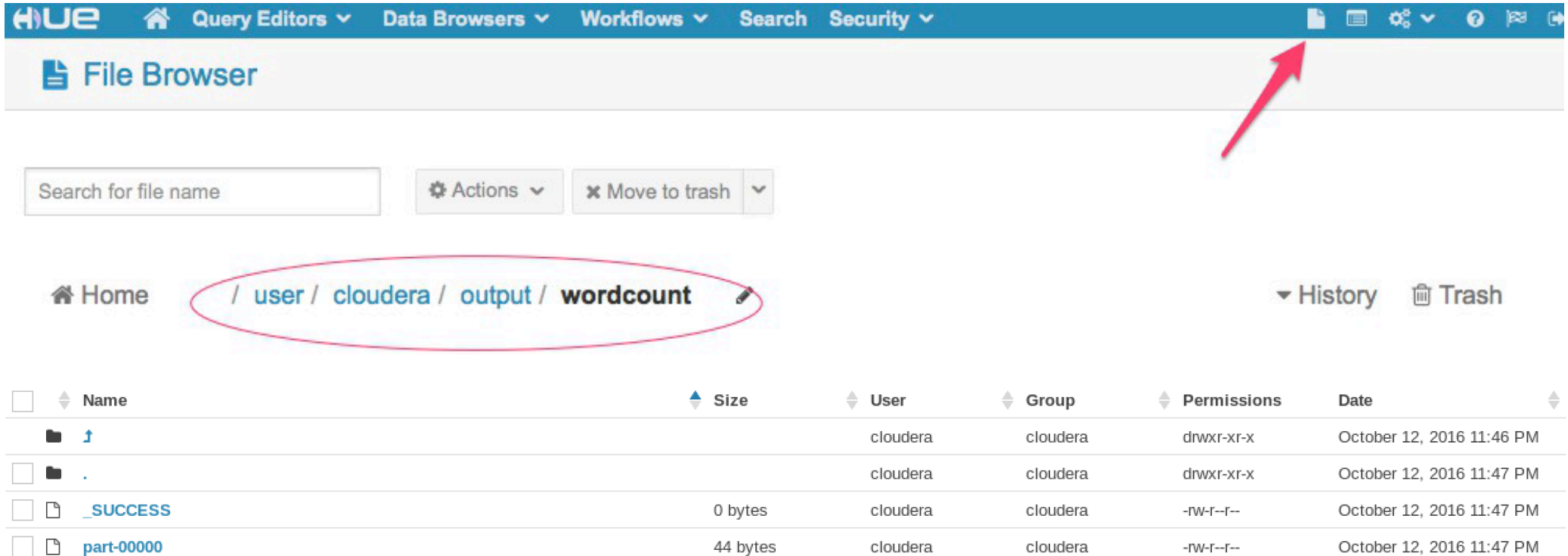
Counters

Recent Tasks





[View All Tasks »](#)

Logs	Tasks	Type
	task_1476334425555_0001_m_000000	MAP
	task_1476334425555_0001_m_000001	MAP
	task_1476334425555_0001_r_000000	REDUCE

Reviewing MapReduce Output Result



The screenshot shows the Hue File Browser interface. The top navigation bar includes links for Query Editors, Data Browsers, Workflows, Search, and Security. The main header is labeled "File Browser". Below the header, there is a search bar and action buttons like "Actions" and "Move to trash". The breadcrumb path is displayed as "Home / user / cloudera / output / wordcount", with the entire path highlighted by a red oval. To the right of the path are links for "History" and "Trash". Below the breadcrumb, a table lists the files and directories in the current location.

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	 ↑		cloudera	cloudera	drwxr-xr-x	October 12, 2016 11:46 PM
<input type="checkbox"/>	 .		cloudera	cloudera	drwxr-xr-x	October 12, 2016 11:47 PM
<input type="checkbox"/>	 _SUCCESS	0 bytes	cloudera	cloudera	-rw-r--r--	October 12, 2016 11:47 PM
<input type="checkbox"/>	 part-00000	44 bytes	cloudera	cloudera	-rw-r--r--	October 12, 2016 11:47 PM

Reviewing MapReduce Output Result

HUE [Home](#) [Query Editors](#) [Data Browsers](#) [Workflows](#) [Search](#) [Security](#)

File Browser

ACTIONS

- View as binary
- Edit file
- Download
- View file location
- Refresh

INFO

Home

Page 1 of 1

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

a	205807
e	315232
i	174282
o	192879
u	65433