CHULALONGKORN
BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Big Data Processing

MapReduce

Thaveewat Khanan

2603615 Organization Data Management

Before MapReduce...

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

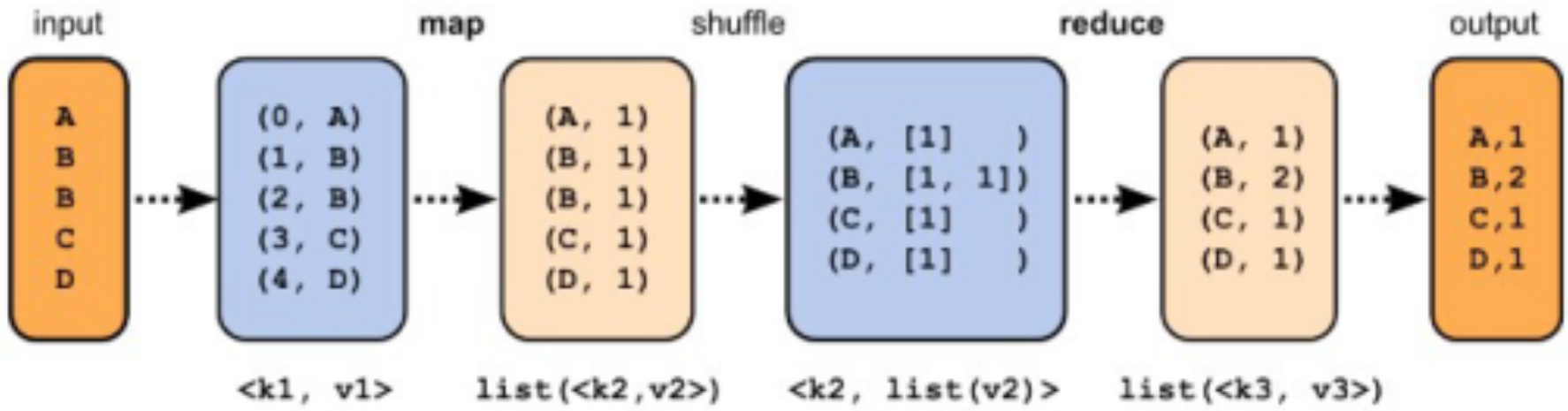
Large scale data processing was difficult!

- 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

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

$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

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

HUE | 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 | Query Editors | Data Browsers | Workflows | Search | Security

Job Browser

JOB ID
1476334425555_0001

TYPE
MR2

USER
cloudera

STATUS
SUCCEEDED

wordcount

Attempts | Tasks | Metadata | Counters

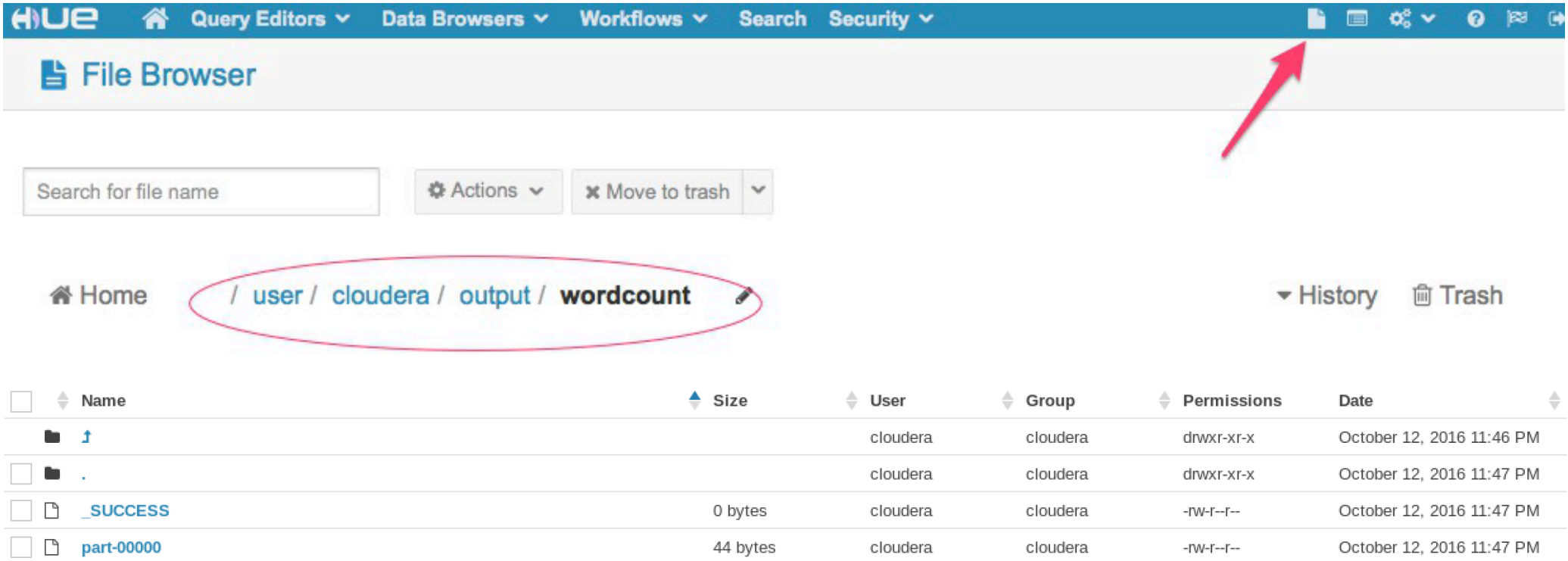
Recent Tasks

View All Tasks »

Logs	Tasks	Type
	task_1476334425555_0001_m_000000	MAP
	task_1476334425555_0001_m_000001	MAP
	task_1476334425555_0001_r_000000	REDUCE

Thaveewat Khanan
2603615 Organization Data Management

Reviewing MapReduce Output Result



Reviewing MapReduce Output Result