

MapReduce 1.0

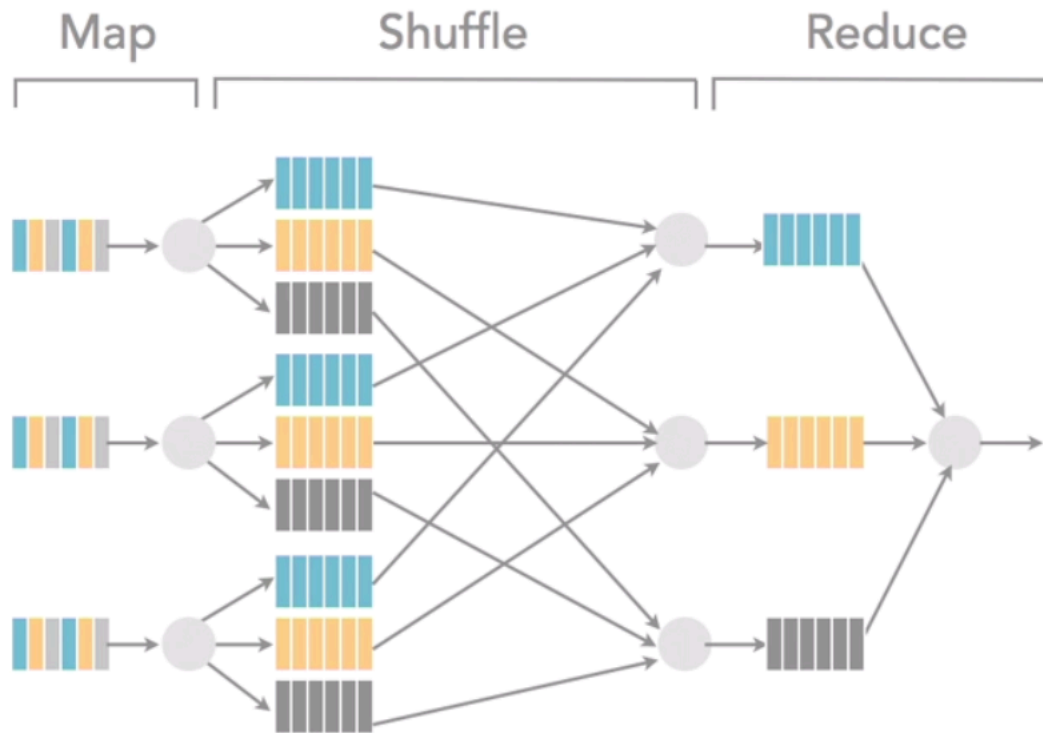
Jay Urbain, PhD

MapReduce

- Programming paradigm
- Originally created at Google
 - *MapReduce: Simplified Data Processing on Large Clusters*, Jeffrey Dean and Sanjay Ghemawat
 - <https://research.google.com/archive/mapreduce.html>
- Designed to solve a single problem: *how to index the Internet!*
- Basically two parts:
 - Map
 - Reduce

MapReduce

- Map(key, value)
 - Execute the Map(key, value) function on data
 - Execute on each node
 - If node goes, its restarted on another node
 - Bring computation to data *versus* bringing the data to the computation.
 - Output(key', value'*) pairs on each node
- Reduce(key', value'*)
 - Execute the Reduce() function to *aggregate* data output from mapper
 - Executes on some of the nodes
 - Output(key'', value''*) combined list



Notes:

- Programmer provides Map and Reduce functions
- Data is sorted as part of shuffle step to reducer
- Shuffle step provided by the framework
- Colors show different *types* of data
- Shows 3 Map nodes on 3 separate physical servers
- Does not show x3 replication

MapReduce 1.0

- Relatively easy to understand
- Distributed, scalable, inexpensive – runs on commodity hardware
- Lends itself to parallel processing, runs on each node, no shared data
- MapReduce 2.0/YARN builds on MapReduce 1.0
- Use MRv2 for new development
- Storage
 - HDFS – triple replicated
- Commodity hardware
- Processing
 - Parallel via Map (local) and Reduce (aggregated)

Coding Steps

- Create a MapReduce class
- Create a static Map function
 - Transform function
- Create a static Reduce function
 - Aggregation function
- Create a main() function
 - Create a job
 - Job calls the Map and Reduce classes
- Programming paradigm:
 - Functional programming: data/state is not shared, data in, data out
 - The equivalent of a complex SQL query has to be broken into multiple steps/jobs/applications

```
public class MapReduce {  
    public static void Main(String[] args)  
    {  
        //create JobRunnerInstance  
        //call MapInstance on JobInstance  
        //call ReduceInstance on JobInstance  
    }  
    public void Map()  
    {  
        //write Mapper  
    }  
    public void Reduce()  
    {  
        //write Reducer  
    }  
}
```

Word count: Hello World for MapReduce

- Google needed to count the number of words on the web for indexing

- Example:
 - How much wood could a woodchuck chuck if a woodchuck could chuck wood?
- Input:
 - List of words (example above)
- Result:
 - {how, 1; much:1, wood,2; could,2; a;2; woodchuck,2; chuck,2; if,1}
- Question:
 - What is a word? Case sensitivity, punctuation, stemming, etc.

Word Count Pseudo Code

Need to look at code with respect to correctness and performance

```
mapper(filename, file-contents)
  for each word in file-contents
    emit(word, 1)
```

```
reducer(word, values)
  sum = 0
  for each value in values:
    sum += value
  emit(word, sum)
```

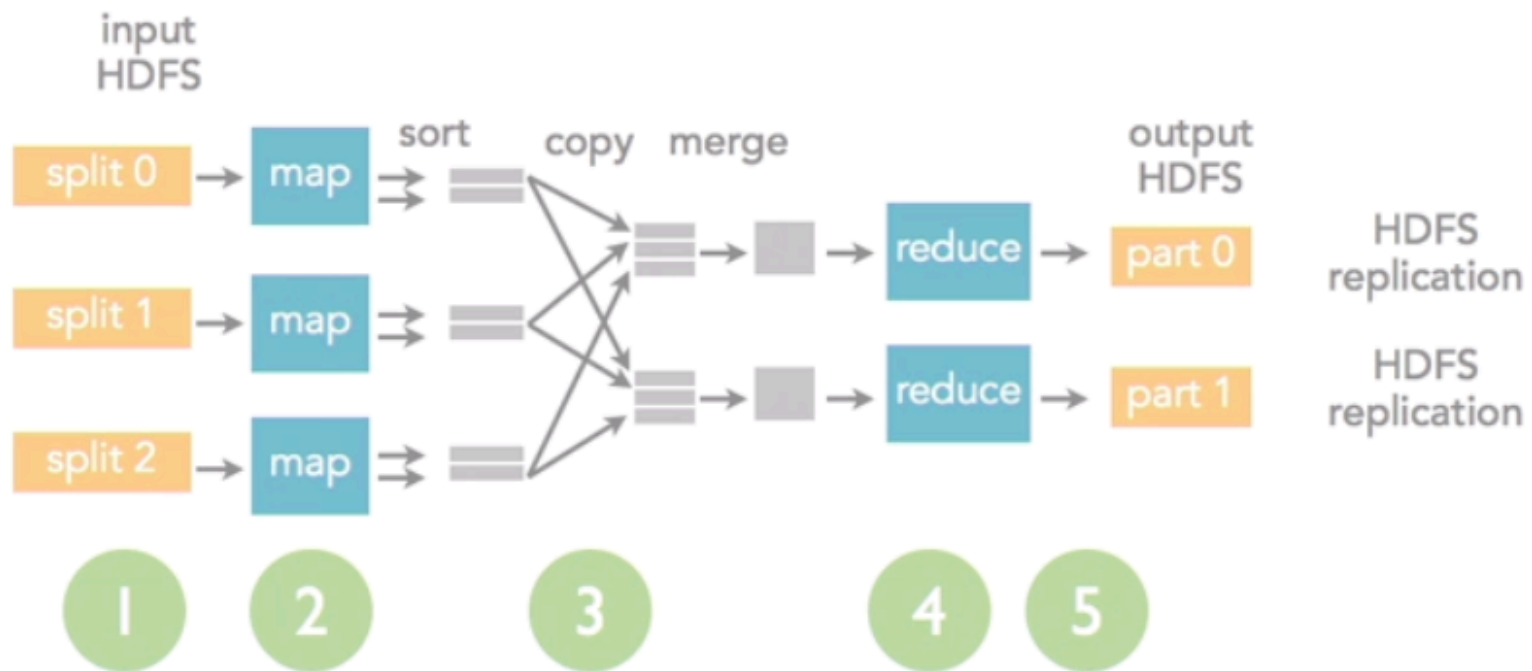

Key Aspects of MapReduce

- MapReduce is an API, set of libraries (jar files)
 - Job – unit of MapReduce work/instance
 - Map task – runs on each node
 - Reduce task – runs on some nodes
 - Source data – HDFS or other location

MapReduce Daemons and Services

- JVMs or services – isolated processes, no shared state
 - Job tracker – one (controller and scheduler)
 - Task trackers – one per cluster host (monitors tasks)
- Job configurations
 - Specify input/output locations for job instances
 - Job clients submit jobs for execution – Use GUI, command line, etc.

MapReduce



- Sort and merge have defaults, but can be overridden
- Input split, number of mappers and reducers are configurable

MapReduce Coding Patterns

- Standard – usually written in Java
- Hadoop Streaming – Java base
 - Other language for mapper/reducer logic
 - Python popular, can use almost anything

Running MapReduce Job

- Word count == hello world
- Can run in IDE, tool, or command line
- Eclipse and Hadoop SDKs – included with Cloudera and Hortonworks distribution
- File system or HDFS
- MapReduce output: SUCCESS, series of delimited text files. Note:
 - Immutable if stored in HDFS
 - Will not allow you to overwrite or update existing files
 - Each run needs a new file name or old files to be removed
- Usually appends run time to file name for uniqueness

MapReduce Job Status and Logs

- Monitor job run status
 - Command line
 - Tools in vendor distribution
 - Log files
 - Troubleshooting failed jobs requires programming and admin skills
 - Error logs – much more verbose than RDBMS systems, can adjust level of logging

Linux Shell Commands

Command	Description
<code>ls</code>	list folder contents
<code>cat</code>	reads (displays) a file
<code>mkdir</code>	makes a directory
<code>cd</code>	change to a directory
<code>sudo command</code>	run <i>command</i> as administrator
<code>chmod file</code>	show/change permissions of <i>file</i>

Hint: Use tab to complete commands

Hadoop Shell Commands

```
hadoop fs -cat file:///file2
```

```
hadoop fs -mkdir /user/hadoop/dir1 /user/hadoop/dir2
```

```
hadoop fs -copyFromLocal <fromDir> <toDir>
```

```
hadoop fs -put <localfile> hdfs://nn.example.com/hadoop/hadoopfile
```

```
sudo hadoop jar <jarFileName> <method> <fromDir> <toDir>
```

```
hadoop fs -ls /user/hadoop/dir1
```

```
hadoop fs -cat hdfs://nn1.example.com/file1
```

```
hadoop fs -get /user/hadoop/file <localfile>
```

Notes:

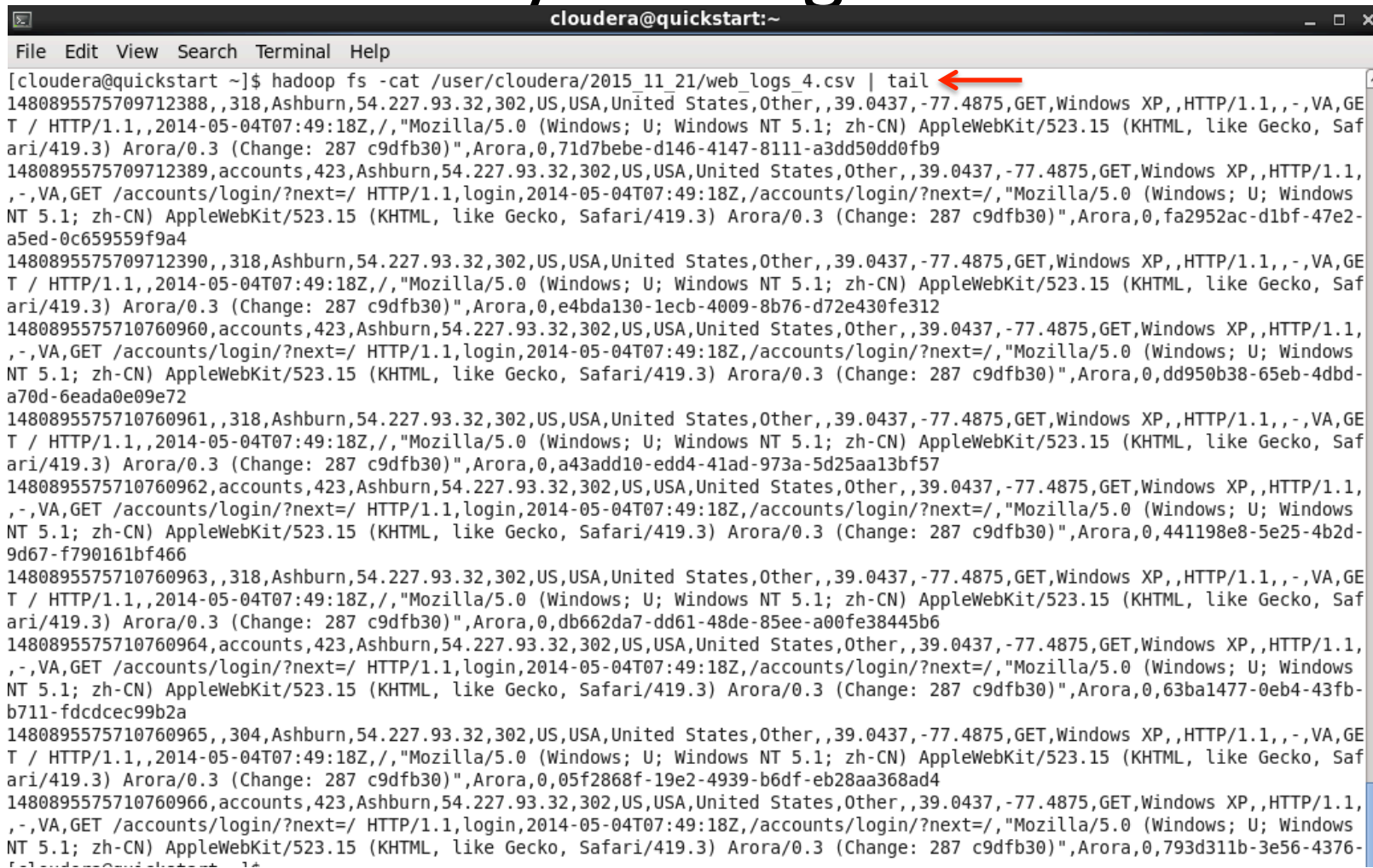
- Some installations use dfs rather than fs
- Old style uses hdfs rather than hadoop

Cloudera VM Terminal

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
[cloudera@quickstart ~]$  
[cloudera@quickstart ~]$  
[cloudera@quickstart ~]$  
[cloudera@quickstart ~]$  
[cloudera@quickstart ~]$  
[cloudera@quickstart ~]$  
[cloudera@quickstart ~]$  
[cloudera@quickstart ~]$  
[cloudera@quickstart ~]$  
[cloudera@quickstart ~]$  
[cloudera@quickstart ~]$  
[cloudera@quickstart ~]$ ls  
cloudera-manager Downloads kerberos Pictures workspace  
cm_api.py eclipse lib Public  
Desktop enterprise-deployment.json Music Templates  
Documents express-deployment.json parcels Videos  
[cloudera@quickstart ~]$ hadoop fs -ls /  
Found 6 items  
drwxrwxrwx - hdfs supergroup 0 2017-04-05 04:27 /benchmarks  
drwxr-xr-x - hbase supergroup 0 2017-06-20 07:16 /hbase  
drwxr-xr-x - solr solr 0 2017-06-20 06:45 /solr  
drwxrwxrwt - hdfs supergroup 0 2017-06-20 07:32 /tmp  
drwxr-xr-x - hdfs supergroup 0 2017-06-20 06:32 /user  
drwxr-xr-x - hdfs supergroup 0 2017-04-05 04:29 /var  
[cloudera@quickstart ~]$
```

Cloudera VM Terminal: head|tail

Handy for huge files

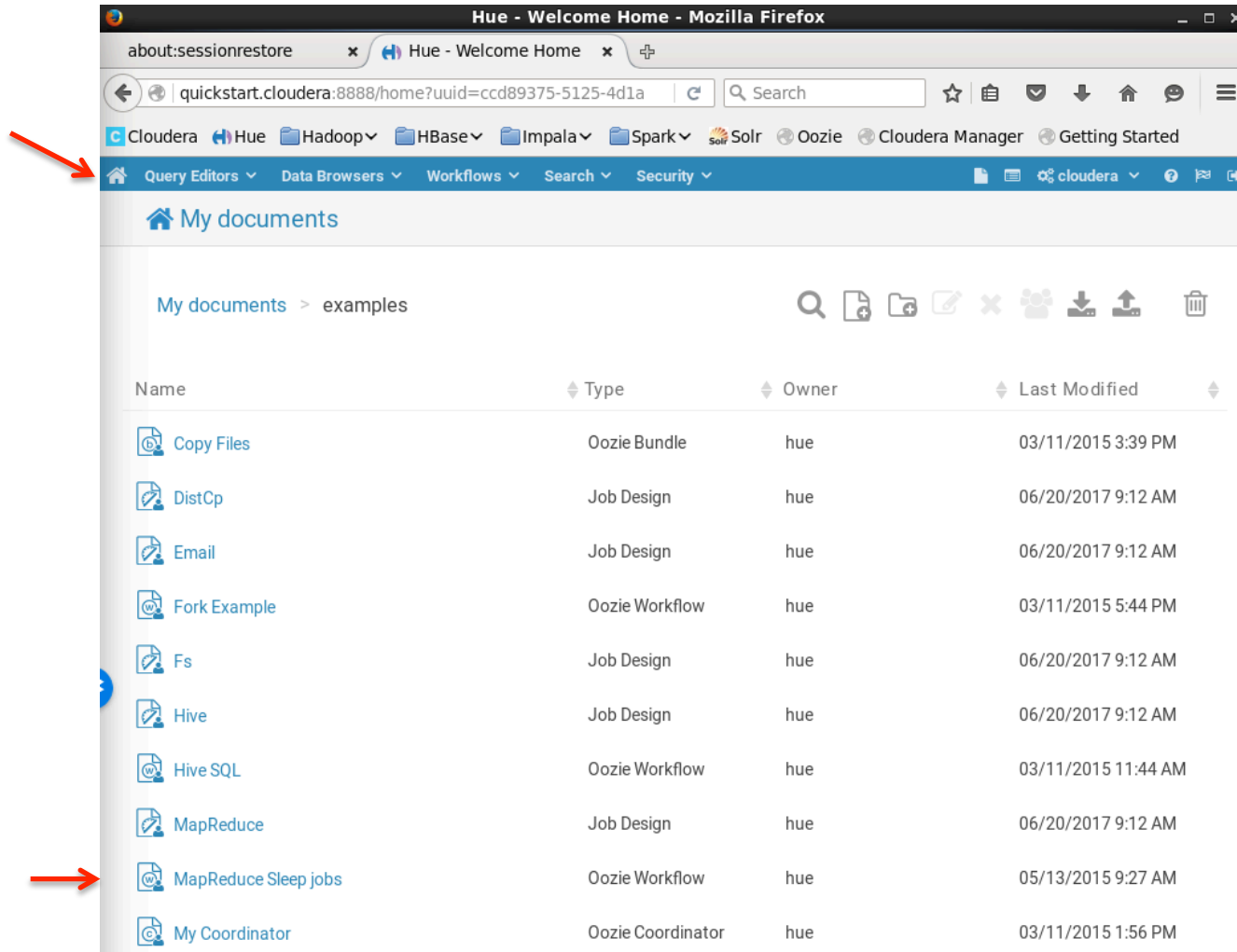


The screenshot shows a terminal window titled "cloudera@quickstart:~". The menu bar includes "File", "Edit", "View", "Search", "Terminal", and "Help". The command prompt shows the user running the command: `[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/2015_11_21/web_logs_4.csv | tail`. A red arrow points to the `tail` command. The terminal output displays a large volume of log data, including IP addresses, user agents, and timestamps, demonstrating the use of `tail` to view the end of a large file.

```
[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/2015_11_21/web_logs_4.csv | tail
1480895575709712388,,318,Ashburn,54.227.93.32,302,US,USA,United States,Other,,39.0437,-77.4875,GET,Windows XP,,HTTP/1.1,,-,VA,GET / HTTP/1.1,,2014-05-04T07:49:18Z,,,"Mozilla/5.0 (Windows; U; Windows NT 5.1; zh-CN) AppleWebKit/523.15 (KHTML, like Gecko, Safari/419.3) Arora/0.3 (Change: 287 c9dfb30)",Arora,0,71d7bebe-d146-4147-8111-a3dd50dd0fb9
1480895575709712389,accounts,423,Ashburn,54.227.93.32,302,US,USA,United States,Other,,39.0437,-77.4875,GET,Windows XP,,HTTP/1.1,,-,VA,GET /accounts/login/?next=/ HTTP/1.1,login,2014-05-04T07:49:18Z,/accounts/login/?next=/"Mozilla/5.0 (Windows; U; Windows NT 5.1; zh-CN) AppleWebKit/523.15 (KHTML, like Gecko, Safari/419.3) Arora/0.3 (Change: 287 c9dfb30)",Arora,0,fa2952ac-d1bf-47e2-a5ed-0c659559f9a4
1480895575709712390,,318,Ashburn,54.227.93.32,302,US,USA,United States,Other,,39.0437,-77.4875,GET,Windows XP,,HTTP/1.1,,-,VA,GET / HTTP/1.1,,2014-05-04T07:49:18Z,,,"Mozilla/5.0 (Windows; U; Windows NT 5.1; zh-CN) AppleWebKit/523.15 (KHTML, like Gecko, Safari/419.3) Arora/0.3 (Change: 287 c9dfb30)",Arora,0,e4bda130-lecb-4009-8b76-d72e430fe312
1480895575710760960,accounts,423,Ashburn,54.227.93.32,302,US,USA,United States,Other,,39.0437,-77.4875,GET,Windows XP,,HTTP/1.1,,-,VA,GET /accounts/login/?next=/ HTTP/1.1,login,2014-05-04T07:49:18Z,/accounts/login/?next=/"Mozilla/5.0 (Windows; U; Windows NT 5.1; zh-CN) AppleWebKit/523.15 (KHTML, like Gecko, Safari/419.3) Arora/0.3 (Change: 287 c9dfb30)",Arora,0,dd950b38-65eb-4dbd-a70d-6eada0e09e72
1480895575710760961,,318,Ashburn,54.227.93.32,302,US,USA,United States,Other,,39.0437,-77.4875,GET,Windows XP,,HTTP/1.1,,-,VA,GET / HTTP/1.1,,2014-05-04T07:49:18Z,,,"Mozilla/5.0 (Windows; U; Windows NT 5.1; zh-CN) AppleWebKit/523.15 (KHTML, like Gecko, Safari/419.3) Arora/0.3 (Change: 287 c9dfb30)",Arora,0,a43add10-edd4-41ad-973a-5d25aa13bf57
1480895575710760962,accounts,423,Ashburn,54.227.93.32,302,US,USA,United States,Other,,39.0437,-77.4875,GET,Windows XP,,HTTP/1.1,,-,VA,GET /accounts/login/?next=/ HTTP/1.1,login,2014-05-04T07:49:18Z,/accounts/login/?next=/"Mozilla/5.0 (Windows; U; Windows NT 5.1; zh-CN) AppleWebKit/523.15 (KHTML, like Gecko, Safari/419.3) Arora/0.3 (Change: 287 c9dfb30)",Arora,0,441198e8-5e25-4b2d-9d67-f790161bf466
1480895575710760963,,318,Ashburn,54.227.93.32,302,US,USA,United States,Other,,39.0437,-77.4875,GET,Windows XP,,HTTP/1.1,,-,VA,GET / HTTP/1.1,,2014-05-04T07:49:18Z,,,"Mozilla/5.0 (Windows; U; Windows NT 5.1; zh-CN) AppleWebKit/523.15 (KHTML, like Gecko, Safari/419.3) Arora/0.3 (Change: 287 c9dfb30)",Arora,0,db662da7-dd61-48de-85ee-a00fe38445b6
1480895575710760964,accounts,423,Ashburn,54.227.93.32,302,US,USA,United States,Other,,39.0437,-77.4875,GET,Windows XP,,HTTP/1.1,,-,VA,GET /accounts/login/?next=/ HTTP/1.1,login,2014-05-04T07:49:18Z,/accounts/login/?next=/"Mozilla/5.0 (Windows; U; Windows NT 5.1; zh-CN) AppleWebKit/523.15 (KHTML, like Gecko, Safari/419.3) Arora/0.3 (Change: 287 c9dfb30)",Arora,0,63ba1477-0eb4-43fb-b711-fdcdcec99b2a
1480895575710760965,,304,Ashburn,54.227.93.32,302,US,USA,United States,Other,,39.0437,-77.4875,GET,Windows XP,,HTTP/1.1,,-,VA,GET / HTTP/1.1,,2014-05-04T07:49:18Z,,,"Mozilla/5.0 (Windows; U; Windows NT 5.1; zh-CN) AppleWebKit/523.15 (KHTML, like Gecko, Safari/419.3) Arora/0.3 (Change: 287 c9dfb30)",Arora,0,05f2868f-19e2-4939-b6df-eb28aa368ad4
1480895575710760966,accounts,423,Ashburn,54.227.93.32,302,US,USA,United States,Other,,39.0437,-77.4875,GET,Windows XP,,HTTP/1.1,,-,VA,GET /accounts/login/?next=/ HTTP/1.1,login,2014-05-04T07:49:18Z,/accounts/login/?next=/"Mozilla/5.0 (Windows; U; Windows NT 5.1; zh-CN) AppleWebKit/523.15 (KHTML, like Gecko, Safari/419.3) Arora/0.3 (Change: 287 c9dfb30)",Arora,0,793d311b-3e56-4376-
```

Hue Home – jobs

Select MapReduce Sleep Jobs



The screenshot shows the Hue Home interface in a Mozilla Firefox browser. The browser address bar displays the URL `quickstart.cloudera:8888/home?uuid=ccd89375-5125-4d1a`. The top navigation bar includes links for Cloudera, Hue, Hadoop, HBase, Impala, Spark, Solr, Oozie, Cloudera Manager, and Getting Started. Below this, a secondary navigation bar contains links for Query Editors, Data Browsers, Workflows, Search, and Security. The main content area is titled 'My documents' and shows a list of jobs under the 'examples' sub-header. The jobs are listed in a table with columns for Name, Type, Owner, and Last Modified. A red arrow points to the 'Query Editors' link in the top navigation bar, and another red arrow points to the 'MapReduce Sleep jobs' entry in the list.

Name	Type	Owner	Last Modified
Copy Files	Oozie Bundle	hue	03/11/2015 3:39 PM
DistCp	Job Design	hue	06/20/2017 9:12 AM
Email	Job Design	hue	06/20/2017 9:12 AM
Fork Example	Oozie Workflow	hue	03/11/2015 5:44 PM
Fs	Job Design	hue	06/20/2017 9:12 AM
Hive	Job Design	hue	06/20/2017 9:12 AM
Hive SQL	Oozie Workflow	hue	03/11/2015 11:44 AM
MapReduce	Job Design	hue	06/20/2017 9:12 AM
MapReduce Sleep jobs	Oozie Workflow	hue	05/13/2015 9:27 AM
My Coordinator	Oozie Coordinator	hue	03/11/2015 1:56 PM

Oozie Workflow

Hue - Workflow Editor - Mozilla Firefox

about:sessionrestore x Hue - Workflow Editor x Hue - Workflow Editor x

quickstart.cloudera:8888/oozie/editor/workflow/edit?workfl Search

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

Query Editors Data Browsers Workflows Search Security

Oozie Editor Workflows Coordinators Bundles

MapReduce Sleep jobs

Run a MapReduce job that sleeps for N seconds

```
graph TD; Start(( )) --> MRJob[MapReduce job  
hadoop-examples.jar]; MRJob --> End(( )); Empty[ ]
```



Application Counts words then sleeps

×

Submit MapReduce Sleep jobs?

REDUCER_SLEEP_TIME

..

☐ Do a dryrun before submitting the job

Cancel

Submit

33% - Hue - Oozie Editor/Dashboard - Workflow Dashboard - Mozilla Firefox

about:sessionrestore x Hue - Workflow Editor x 33% - Hue - Oozie Ed... x

quickstart.cloudera:8888/oozie/list_oozie_workflow/000000 Search

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

Query Editors Data Browsers Workflows Search Security cloudera

Oozie Dashboard Workflows Coordinators Bundles SLA Oozie

WORKFLOW

Workflow MapReduce Sleep jobs

MapReduce Sleep jobs

Graph Actions Details Configuration Log Definition

SUBMITTER

cloudera

STATUS

RUNNING

PROGRESS

33%

ID

0000000-170620071
oozie-oozi-W

VARIABLES

send_email

REDUCER_SLEEP_TIM

submit_single_action **Back**

oozie.wf.ap...

MANAGE

Kill

Suspend

All jobs
Have an ID

Edit Properties

The screenshot shows the Hue Workflow Editor interface in Mozilla Firefox. The browser address bar displays the URL: `quickstart.cloudera:8888/oozie/editor/workflow/edit/?workfl`. The interface includes a top navigation bar with tabs for Query Editors, Data Browsers, Workflows, Search, and Security. Below this is a toolbar with various icons for workflow management. The main workspace is titled "Oozie Editor" and "Workflows". A workflow diagram is visible, showing a sequence of steps. The first step is a "MapReduce job" configuration panel. This panel includes a "Jar name" field with the value `/user/hue/oozie/workspaces/lib/h...`. Below the jar name is a "PROPERTIES" section with a table of configuration properties.

Property	Value
<code>.reduce.tasks</code>	<code>1</code>
<code>mapred.mapr...</code>	<code>org.apache.hadoop.exam...</code>
<code>mapred.reduc...</code>	<code>org.apache.hadoop.exam...</code>
<code>mapred.mapc...</code>	<code>org.apache.hadoop.io.Int...</code>
<code>mapred.mapc...</code>	<code>org.apache.hadoop.io.Nul...</code>
<code>mapred.outpu...</code>	<code>org.apache.hadoop.mapr...</code>
<code>mapred.input...</code>	<code>org.apache.hadoop.exam...</code>
<code>mapred.partit...</code>	<code>org.apache.hadoop.exam...</code>

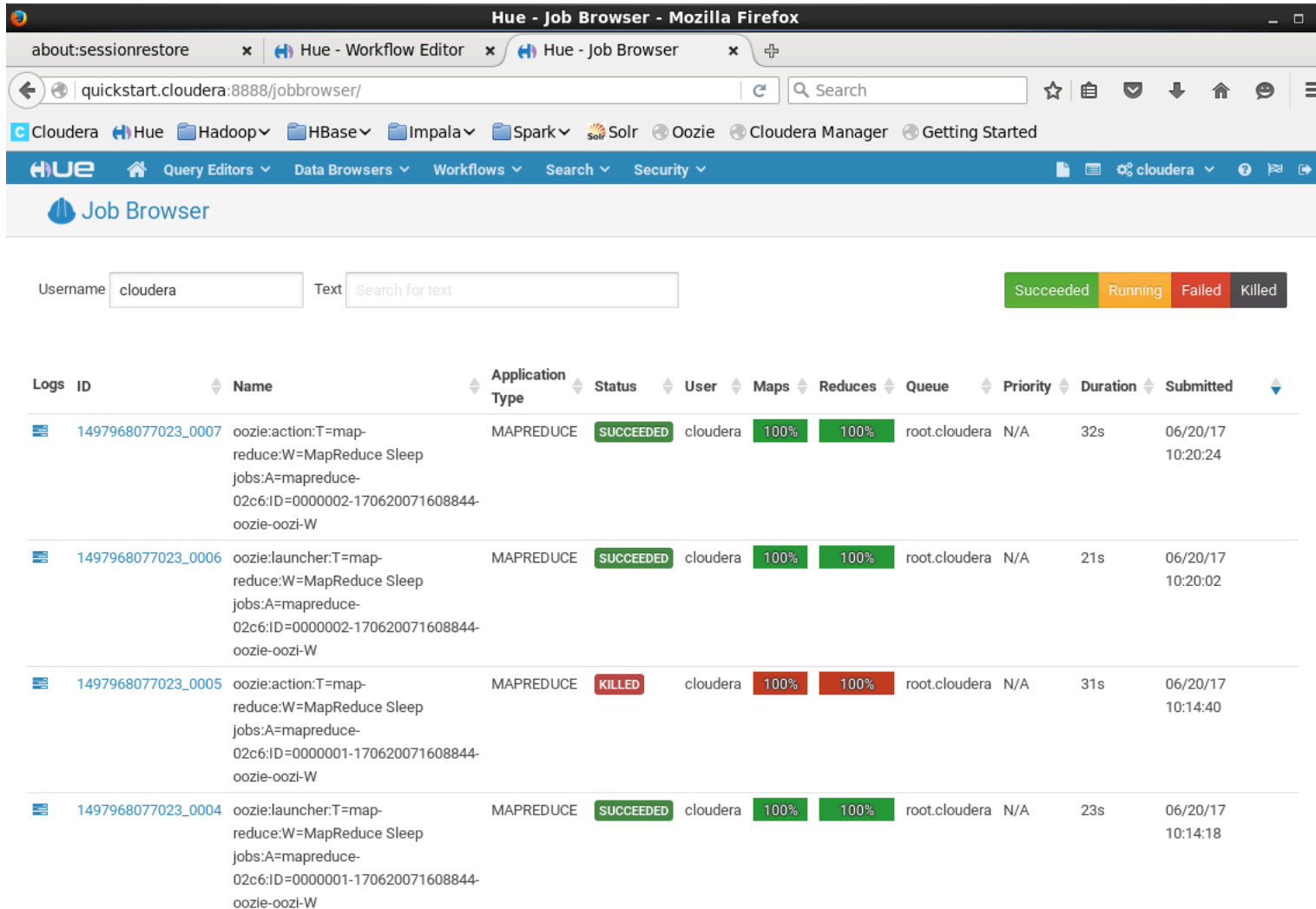
Oozie Dashboard – Job Status

The screenshot shows the Oozie Dashboard in a Mozilla Firefox browser. The address bar displays the URL `quickstart.cloudera:8888/oozie/list_oozie_workflows/`. The browser's navigation bar includes links for Cloudera, Hue, Hadoop, HBase, Impala, Spark, Solr, Oozie, Cloudera Manager, and Getting Started. The Oozie Dashboard header features tabs for Query Editors, Data Browsers, Workflows (selected), Search, and Security. Below the header, the 'Workflows' section is active, showing a search bar and filters for Running, Succeeded, Error, Manually, and Coordinator. The 'Running' section is currently empty, displaying 'No matching records' and 'Showing 0 to 0 of 0 entries'. The 'Completed' section shows a table of finished jobs.

Completion	Status	Name	Duration	Submitter	Id	Parent
Tue, 20 Jun 2017 10:20:58	SUCCEEDED	MapReduce Sleep jobs	56s	cloudera	0000002-170620071608844-oozie-oozi-W	
Tue, 20 Jun 2017 10:15:11	KILLED	MapReduce Sleep jobs	54s	cloudera	0000001-170620071608844-oozie-oozi-W	
Tue, 20 Jun 2017 09:25:07	SUCCEEDED	MapReduce Sleep jobs	59s	cloudera	0000000-170620071608844-oozie-oozi-W	

Showing 1 to 3 of 3

Job Browser (via logs view)



The screenshot shows the Hue Job Browser interface in a Mozilla Firefox browser. The browser tabs include 'about:sessionrestore', 'Hue - Workflow Editor', and 'Hue - Job Browser'. The address bar shows 'quickstart.cloudera:8888/jobbrowser/'. The Hue navigation bar includes links for 'Query Editors', 'Data Browsers', 'Workflows', 'Search', and 'Security'. The 'Job Browser' section has a search bar with 'Username: cloudera' and a 'Text' field. Below the search bar is a table of jobs.

Logs ID	Name	Application Type	Status	User	Maps	Reduces	Queue	Priority	Duration	Submitted
1497968077023_0007	oozie:action:T=map-reduce;W=MapReduce Sleep;jobs:A=mapreduce-02c6;ID=0000002-170620071608844-oozie-oozi-W	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	32s	06/20/17 10:20:24
1497968077023_0006	oozie:launcher:T=map-reduce;W=MapReduce Sleep;jobs:A=mapreduce-02c6;ID=0000002-170620071608844-oozie-oozi-W	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	21s	06/20/17 10:20:02
1497968077023_0005	oozie:action:T=map-reduce;W=MapReduce Sleep;jobs:A=mapreduce-02c6;ID=0000001-170620071608844-oozie-oozi-W	MAPREDUCE	KILLED	cloudera	100%	100%	root.cloudera	N/A	31s	06/20/17 10:14:40
1497968077023_0004	oozie:launcher:T=map-reduce;W=MapReduce Sleep;jobs:A=mapreduce-02c6;ID=0000001-170620071608844-oozie-oozi-W	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	23s	06/20/17 10:14:18

Tasks

Hue - Job Browser - Job: 1497968077023_0007 - Mozilla Firefox

about:sessionrestore x Hue - Workflow Editor x Hue - Job Browser - J... x

quickstart.cloudera:8888/jobbrowser/jobs/application_1497968077023_0007

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

HUE Query Editors Data Browsers Workflows Search Security cloudera

Job Browser

JOB ID

1497968077023_...

TYPE

MR2

USER

cloudera

STATUS

SUCCEEDED

LOGS

Logs

MAPS

100%

REDUCES

100%

DURATION

23s

oozie:action:T=map-reduce:W=MapReduce Sleep jobs:A=mapreduce-02c6:ID=0000002-170620071608844-oozie-oozi-W

Attempts Tasks Metadata Counters

Recent Tasks

[View All Tasks »](#)

Logs	Tasks	Type
	task_1497968077023_0007_m_000000	MAP
	task_1497968077023_0007_m_000001	MAP
	task_1497968077023_0007_r_000000	REDUCE

Design Your Own Project – Hue -> Query Editor -> Job Designer

Hue - Job Designer - Mozilla Firefox

about:sessionrestore x Hue - Workflow Editor x Hue - Job Designer x

quickstart.cloudera:8888/jobsub/#list-designs

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

HUE Query Editors Data Browsers Workflows Search Security cloudera

Job Designer

Designs

Search for design name Submit Edit Copy Move to trash New action View trash

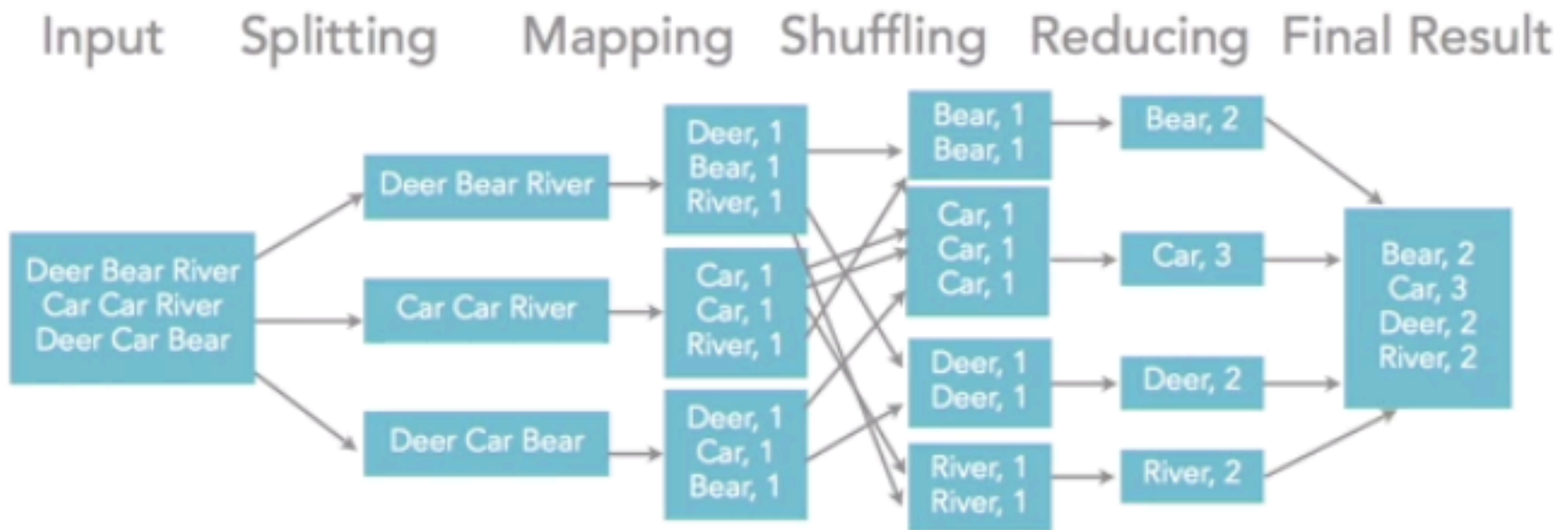
<input type="checkbox"/>	Name	Description	Owner	Type	Status	Last modified
<input type="checkbox"/>	Ssh	Example of SSH action	hue	ssh	shared	June 20, 2017 09:12 AM
<input type="checkbox"/>	Hive	Example of Hive action	hue	hive	shared	June 20, 2017 09:12 AM
<input type="checkbox"/>	DistCp	Example of DistCp action	hue	distcp	shared	June 20, 2017 09:12 AM
<input type="checkbox"/>	Shell	Example of Shell action	hue	shell	shared	June 20, 2017 09:12 AM
<input type="checkbox"/>	Pig	Example of Pig action	hue	pig	shared	June 20, 2017 09:12 AM
<input type="checkbox"/>	Fs	Example of Fs action	hue	fs	shared	June 20, 2017 09:12 AM
<input type="checkbox"/>	MapReduce	Example of MapReduce action that sleeps	hue	mapreduce	shared	June 20, 2017 09:12 AM
<input type="checkbox"/>	Sqoop	Example of Sqoop action	hue	sqoop	shared	June 20, 2017 09:12 AM
<input type="checkbox"/>	Email	Example of Email action	hue	email	shared	June 20, 2017 09:12 AM

Showing 1 to 9 of 9 entries

Previous 1 Next

MapReduce Word Count Example

MapReduce Example - Word Count



Developers focus on map and reduce, can override defaults

MapReduce API Versions

- Version 1.0
 - org.apache.hadoop.**mapred**
- Version 2.0
 - org.apache.hadoop.**mapreduce**

MapReduce Libraries

```
import org.apache.hadoop.fs.Path;  
import org.apache.hadoop.conf.*;  
import org.apache.hadoop.io.*;  
import org.apache.hadoop.mapred.*;  
import org.apache.hadoop.util.*;
```

MapReduce Mapper Code (Version 1)

```
public static class Map extends MapReduceBase
    implements Mapper<LongWritable, Text, Text, IntWritable> {
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(LongWritable key, Text value,
        OutputCollector<Text, IntWritable> output, Reporter reporter)
        throws IOException {
        String line = value.toString();
        StringTokenizer tokenizer = new StringTokenizer(line);
        while (tokenizer.hasMoreTokens()) {
            word.set(tokenizer.nextToken());
            output.collect(word, one);
        }
    }
}
```

MapReduce Reducer Code

```
public static class Reduce extends MapReduceBase
    implements Reducer<Text, IntWritable, Text, IntWritable> {
    public void reduce(Text key, Iterator<IntWritable> values,
        OutputCollector<Text, IntWritable> output, Reporter reporter)
        throws IOException {
        int sum = 0;
        while (values.hasNext()) {
            sum += values.next().get(); // aggregation
        }
        output.collect(key, new IntWritable(sum));
    }
}
```

MapReduce Job Main

```
public static void main(String[] args) throws Exception {  
    JobConf conf = new JobConf(WordCount.class);  
    conf.setJobName("wordcount");  
    conf.setOutputKeyClass(Text.class);  
    conf.setOutputValueClass(IntWritable.class);  
    conf.setMapperClass(Map.class);  
    conf.setCombinerClass(Reduce.class); // adds up values on each node before network shuffle  
    conf.setReducerClass(Reduce.class);  
    conf.setInputFormat(TextInputFormat.class);  
    conf.setOutputFormat(TextOutputFormat.class);  
    FileInputFormat.setInputPaths(conf, new Path(args[0]));  
    FileOutputFormat.setOutputPath(conf, new Path(args[1]));  
    JobClient.runJob(conf);  
}
```


Key Components

- Input/output (data)
 - Writable/write comparable
- Mapper
 - Maps input key/value pairs to a set of intermediate key/value pairs.
- Reducer
 - Reduces a set of intermediate values which share a key to a smaller set of values.
- Partitioner
 - Controls the partitioning of the keys of the intermediate map-outputs.
- Reporter
 - A facility for reporting progress and update counters, status information etc.
- OutputCollector
 - Collect data output by either the Mapper or the Reducer i.e. intermediate outputs or the output of the job.

Writable Data Types

- Can also write custom writable data type

Class	Size (bytes)	Description	Sort Policy
BooleanWritable	1	Wrapper for standard Boolean variable	False before, true after
ByteWritable	1	Wrapper for single byte	Ascending
DoubleWritable	8	Wrapper for a Double	Ascending
FloatWritable	4	Wrapper for a Float	Ascending
IntWritable	4	Wrapper for an Integer	Ascending

Input Types

Format	Description
TextInputFormat	Each line in a text file is a record <LongWritable (offset of line), Text (content of line)>
KeyValueTextInputFormat	Each line is a record. First separator divides line (\t) <Text (before separator), Text (after separator)>
SequenceFileInputFormat<K,V>	Sequence for reading files
NLineInputFormat	Like TextInputFormat; each split has exactly N lines <LongWritable, Text>

Output Types

Format	Description
TextOutputFormat	Write each record as a line of text. Keys and values are written as strings and separated by \t
SequenceFileOutputFormat<K, V>	Write the key/value pairs in sequence file format. Works with SequenceFileInputFormat.
NullOutputFormat<K,V>	Outputs nothing

Running and tracking Hadoop Jobs

- Configure JobConf options
- From the development environment (IDE)
- From a GUI (Hue / HDInsight console)
- From the command line:
 - *hadoop jar filename.jar input output*

Job Execution Optimizations

- Speculative execution
 - kills long-running jobs, and (trys) restarts
 - Based on configuration parameters

Methods to write MapReduce Jobs

- **Standard – usually written in Java**
- Streaming (Python common)
- Pipes (C++ common)
- Abstraction libraries
 - Hive, Pig, et. (higher-level language)
 - Generate MapReduce jobs

Ways to use MapReduce

Need to determine what level of abstraction you want to work at
Java gives you a lot of control

Libraries	Languages
HBase	Java
Hive	HiveQL (HQL)
Pig	Pig Latin
Sqoop	Python
Oozie	C#
Mahout	JavaScript
Others	R

Review

- Which of the following are key aspects of MapReduce Version 1?
 - It is extremely distributed, scalable, and cheap
 - It is resilient because if one of the nodes goes down, HDFS is self-healing
 - It really lends itself to parallel processing
 - All of the answers
- Which of the following terminal commands will read the contents of a file from the local file system?
 - `hadoop fs -cat file:///file2`
 - **`hadoop fs -cat hdfs:///file1`**
 - `hadoop fs -get /user/hadoop/file <localfile>`
 - `hadoop fs -copyFromLocal <fromDir> <toDir>`

Review

- If you're using HDFS in addition to MapReduce API, you're going to have some core libraries that you need. Which of the following is a core library?
 - All of these answers
 - `org.apache.hadoop.io.*`
 - `org.apache.hadoop.conf.*`
 - `org.apache.fs.Path.*`
- Which of the following is not a method to write MapReduce Job?
 - Standard – usually Java
 - Streaming paradigm
 - Pipes
 - JobConf