CS542000 Cloud Programming
Lab3: Hadoop & MapReduce

Josh Kao

National Tsing Hua University

2015/04/27

Objective

- To get familiar with using HDFS & MapReduce
- To customize your own **key**, **value** class

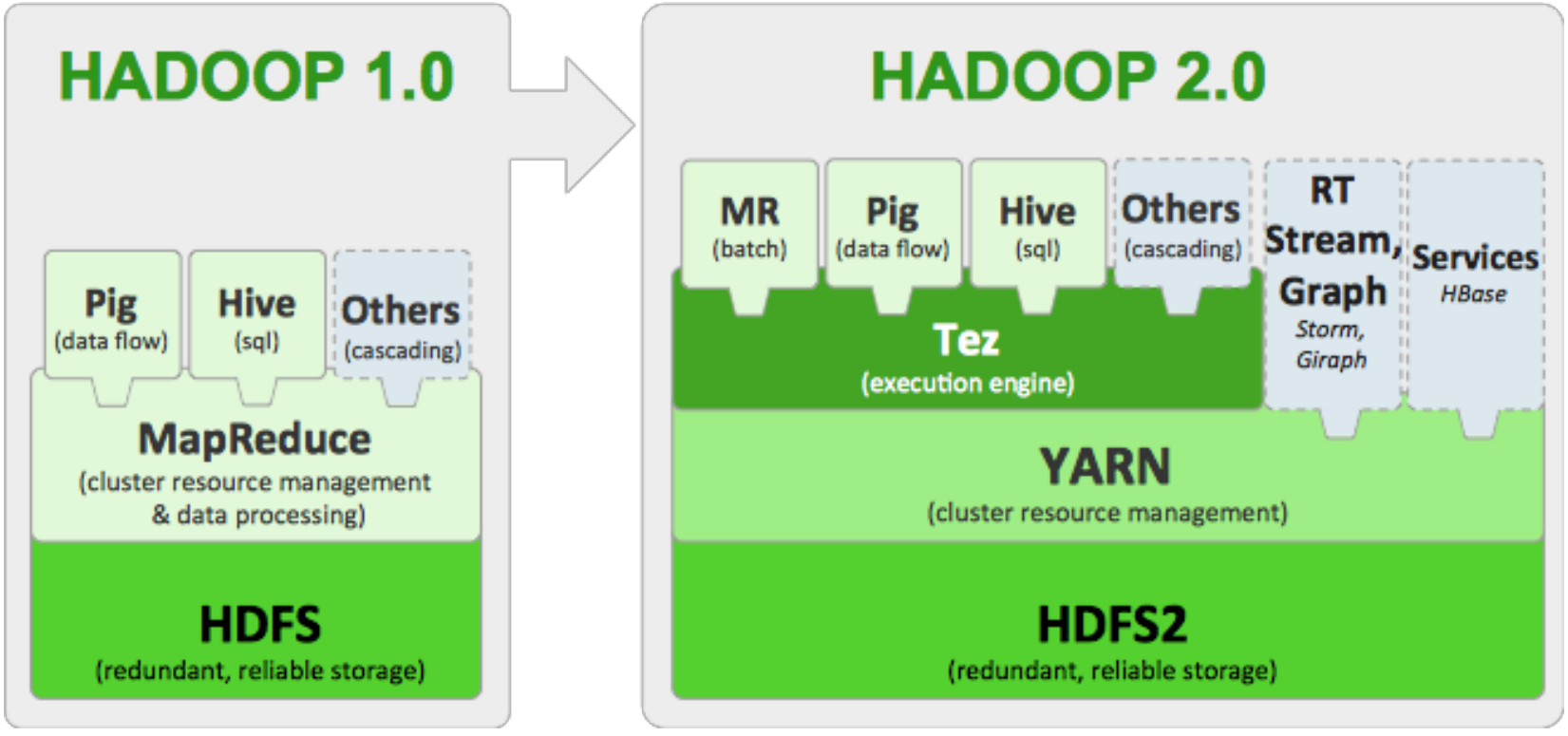
Outline I

What is Hadoop?

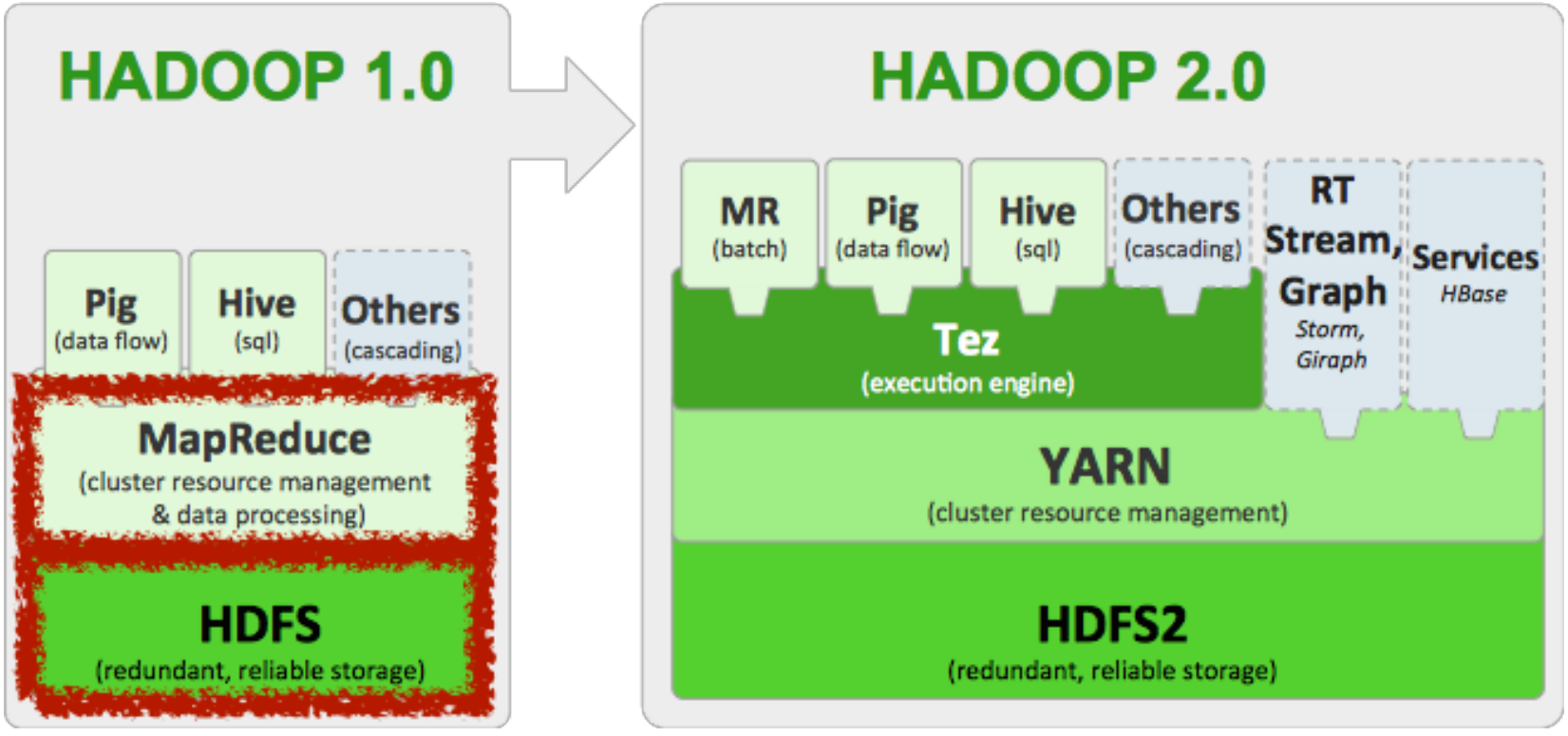
"*Apache Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware.*"

— Wikipedia

Hadoop ecosystem evolution

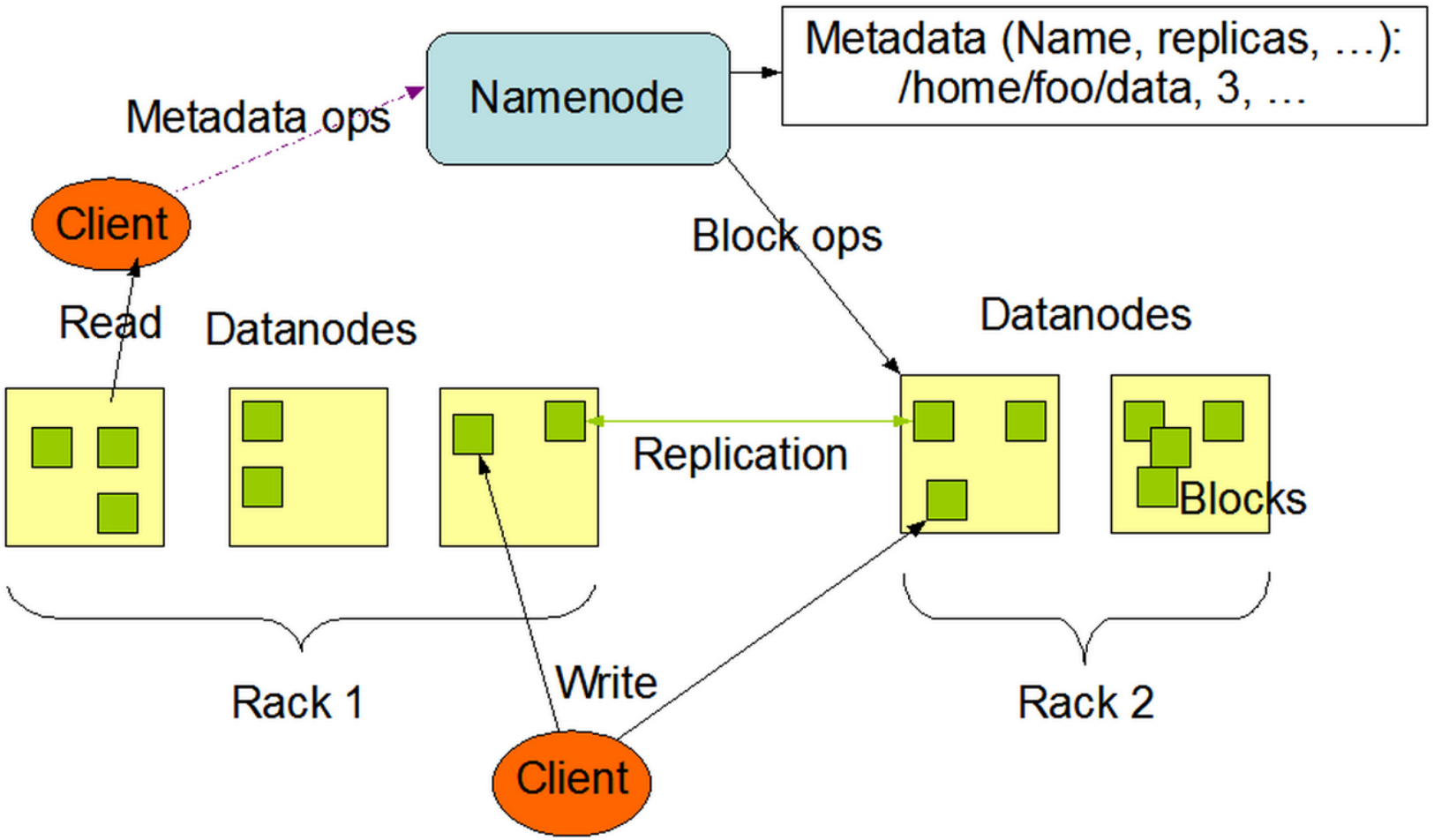


Today's mission



Outline I

Architecture



Commands

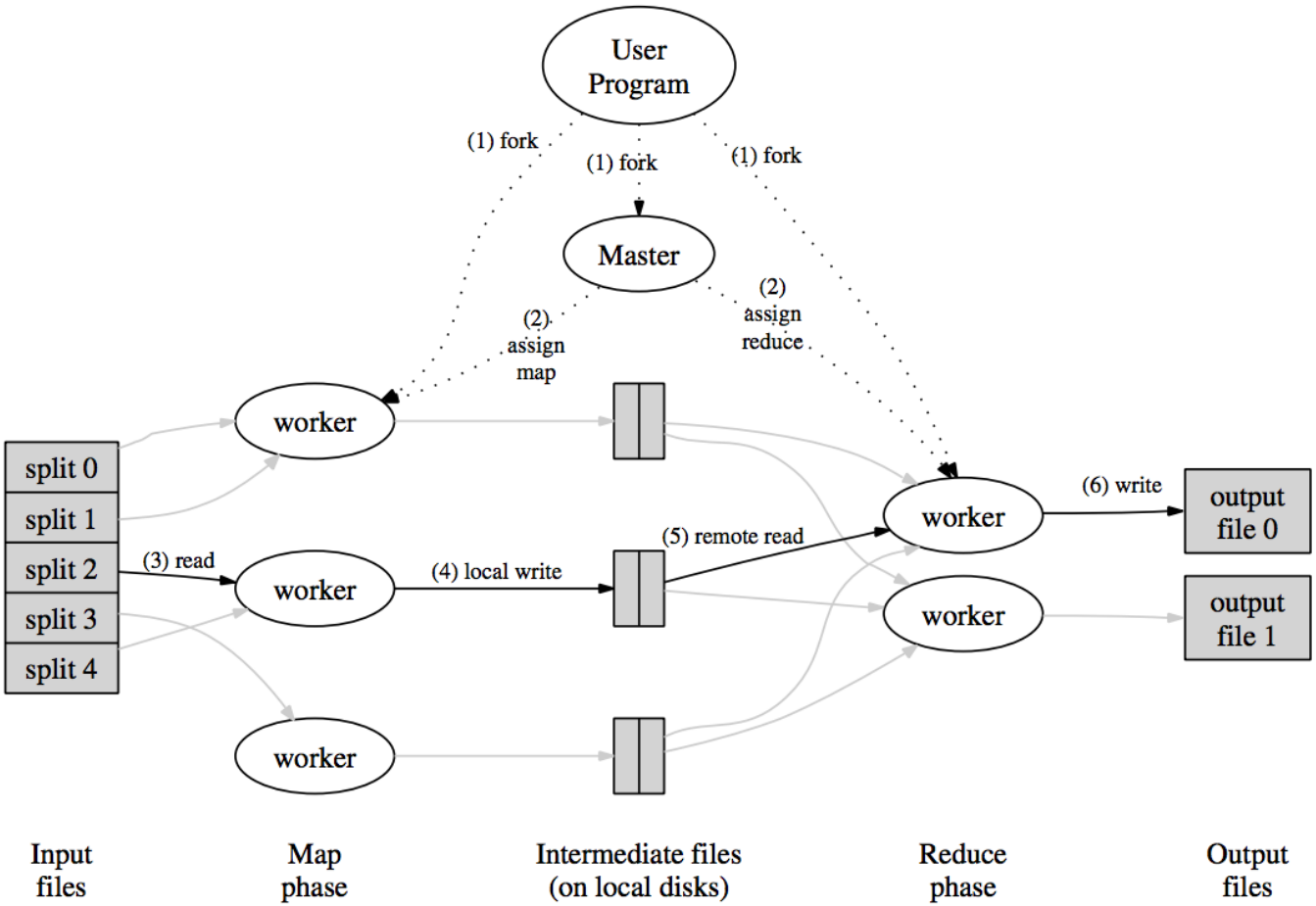
- Make a directory named "*Dir5566*"
  - `$ hadoop dfs -mkidr Dir5566`
- Put file "*IamSad*"(Local) into directory "*Dir5566*"(HDFS)
  - `$ hadoop dfs -put IamSad Dir5566`
  - You can also update directory to HDFS
- List files recursively in directory "*Dir5566*"
  - `$ hadoop dfs -lsr Dir5566`

Commands Cont'

- Check all content out in directory "*Dir5566*"
  - `$ hadoop dfs -cat Dir5566`
- Get directory "*Dir5566*"(HDFS) to local
  - `$ hadoop dfs -get Dir5566 .`
- Remove directory (files) recursively
  - `$ hadoop dfs -rmr Dir5566`

Outline I

MapReduce Workflow



Outline I

Problem Description

- Part 1 - Word Count EX
  'Word Definition Game' is a game to guess word by a given start letter and the meaning of the word. Your job is to
  1. Calculate occurrence rates of each start letter;
  2. Result should be case sensitive;
  3. Cause of limited computing resource, you can use only 2 reducers;
  4. For load balance issue, you have to make first reducer process words start with Aa $\sim$ Gg, and sencond reducer process remaining words.

Word Count EX Sample I/O

Sample Input

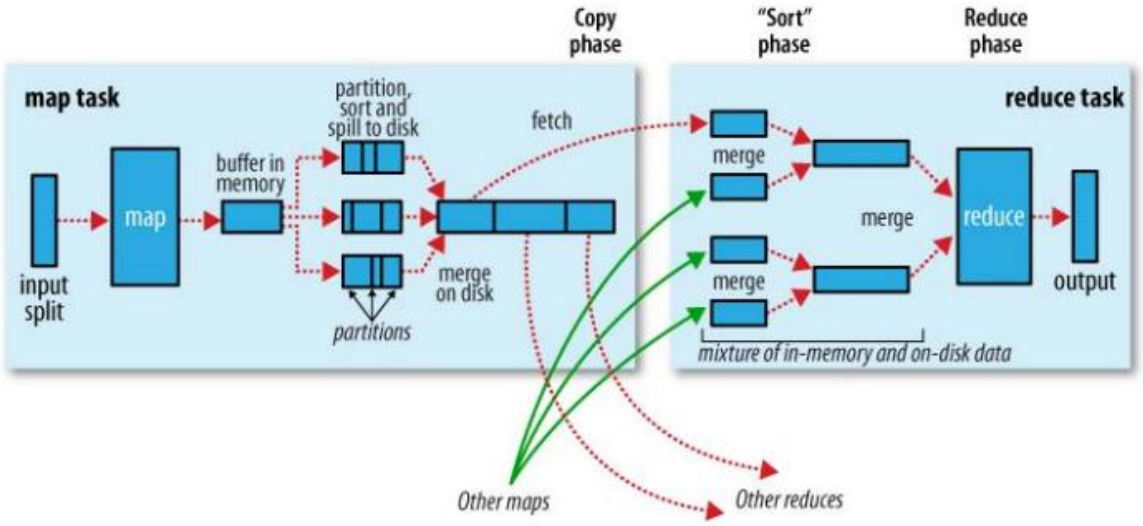
Apache Hadoop is an open-source software framework for storage large-scale processing of data-sets on clusters of commodity hardware.

Sample Output

A : 1
a : 1
B : 0

What's the issue?

- Work Flow



- Component
  - **Partitioner** : Assign specific job to each reducer
  - **Key Comparator** : Sort
  - **Group Comparator** : Group words with the same start letter

Problem Description

- Part 2 - Calculate Average
  The input contains many word-value paris, your job is to
  1. Calculate average value of each word;
  2. Speedup by Combiner

Calcuate Average Sample I/O

Sample Input

cat 5
dog 3
cat 7
dog 10
bear 1

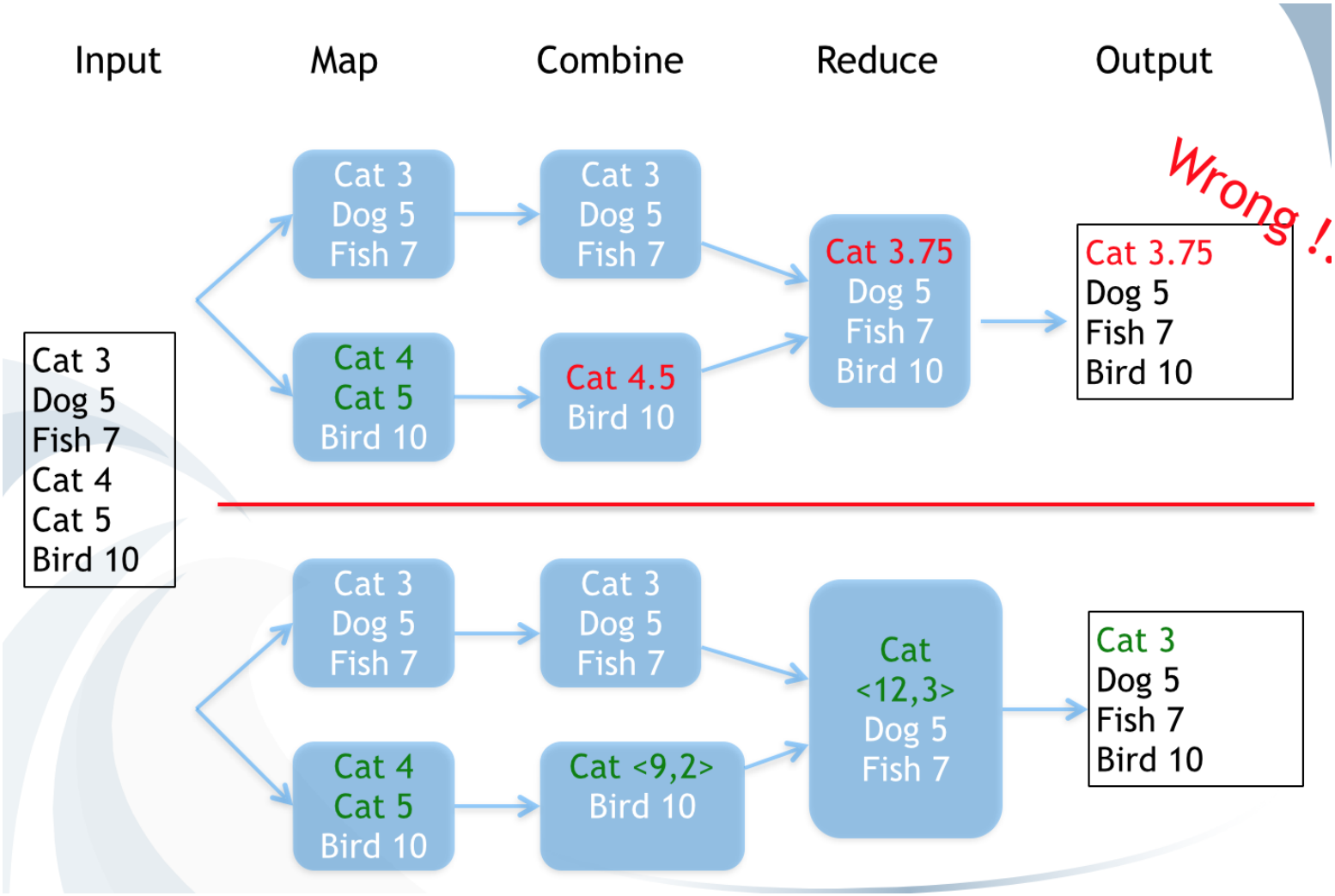
Sample Output

bear 1
cat 6
dog 6.5

What's the issue?

- Component
  - Customize your key-value pair in *MyMapper.java*
  - Record 2 messages each value **(sum, count)**
  - Count average in *MyReducer.java*
  - Speed up with *MyCombiner.java*

What's the issue?



Outline I

Login to server

- **Host**: 140.114.91.199 (with 1 master, 8 slaves)
- **Account**: [Your Student ID]
- **Password**: cloud5566

It's recommended to use ***passwd*** to change your password.

How to compile?

Take wordcount for instance,

1. Compile java source codes to class files
   - javac -classpath *[Depend Jars]* -d *[ClassFolder]* *[Java Files . . .]*
   - `$ javac -classpath hadoop-core-1.0.3.jar -d class/ java/*`
2. Pack class files to a jar file
   - jar -cvf *[Jar Name]* -C *[ClassFolder]* *[Target Directory]*
   - `$ jar -cvf wordcount.jar -C class .`

How to submit a job?

- hadoop jar *[Jar File] [PackageName.ClassName] [ARGS . . .]*
- $ hadoop jar wordcount.jar part1.WordCount input output

Monitoring by Web Interface

- MapReduce Job Tracker
  - http://140.114.91.199:50030
- HDFS
  - http://140.114.91.199:50070

Grading

- Part 1 - Word Count EX
  - 20% Partitioner
  - 20% Key Comparator
  - 20% Group Comparator
- Part 2 - Calculate Average
  - 20% Customize Mapper class
  - 20% Use Combiner to speed up

Outline I

Reference

- http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html
- http://hortonworks.com/blog/apache-hadoop-2-is-ga/
- http://www.bodhtree.com/blog/tag/hadoop/page/4/