

# Apache Hadoop, Big Data, and You

## QCon

---

Philip Zeyliger

philip@cloudera.com

 @philz42 @cloudera

November 18, 2009



# Hi there!

- \* Software Engineer

- \* Worked at  

- \* 

# I work on stuff...

The screenshot displays the Cloudera Desktop interface. On the left, the 'cluster health\_' dashboard shows a grid of nodes with status indicators (green for healthy, red for troubled, grey for critical). Below this is a 'File Browser' window showing a directory structure for '/user/philip/shakespeare' with files like 'a\_and\_c.xml', 'all\_well.xml', 'as\_you.xml', 'com\_err.xml', 'coriolan.xml', and 'cymbelin.xml'. The 'Job Browser' window is open, displaying a table of jobs with columns for Id, Status, User, Name, Maps / Reduces, and Queue. The jobs listed include 'streamjob2979679455937988510.jar', 'streamjob5607796618053275287.jar', and several 'Shakespeare Word Index' jobs. One job (Id 0027) is in a failed state (red exclamation mark). At the bottom, there are buttons for 'new streaming' and 'new jar', and a table showing job details for 'Shakespeare Word Index' and 'Example: Sleep Job' and 'Example: Pi Calculator'.

Cloudera Desktop

Hi philip [logout] Launch ▾

cluster health\_

dn GM GM jt nn tt

File Browser

/user/philip/shakespeare

filter nodes: healthy troubled critical

clusters: GANGLIA(default) 2, HDFS(default) 2, MR(default) 2, Host Summary 1

by role: ann, snn, jt, tt, dn

a\_and\_c.xml  
all\_well.xml  
as\_you.xml  
com\_err.xml  
coriolan.xml  
cymbelin.xml

09/28/2009 8:49am 204.8 KB  
09/28/2009 8:49am 187.6 KB  
09/28/2009 8:49am 133.6 KB  
09/28/2009 8:49am 254 KB  
09/28/2009 8:49am 241.6 KB

Job Browser

Job Browser » List

Filters: All states filter on user text filter

Id	Status	User	Name	Maps / Reduces	Queue
0022	✓	philip	streamjob2979679455937988510.jar	37/37 1/1	default
0024	✓	philip	streamjob5607796618053275287.jar	37/37 1/1	default
0025	✓	philip	Shakespeare Wordcount	37/37 1/1	default
0027	!	philip	Shakespeare Word Index	8/37 0/1	default
0028	✓	philip	Shakespeare Word Index	37/37 1/1	default
0029	✓	philip	Shakespeare Word Index	37/37 1/1	default

history

filter by owner filter by name

Owner	Name	Last Modified
philip	Shakespeare Word Index	7 hours, 58 minutes ago
sample	Example: Sleep Job	2 days, 3 hours ago
sample	Example: Pi Calculator	1 week, 1 day ago

+ new streaming  
+ new jar

# Outline

- \* Why should you care? (Intro)
- \* Challenging yesteryear's assumptions
- \* The MapReduce Model
- \* HDFS, Hadoop Map/Reduce
- \* The Hadoop Ecosystem
- \* Questions

Data is everywhere.

Data is important.

DISCOVER.  
PARTICIPATE.  
ENGAGE.

Search the following Data.gov catalogs:



"RAW" DATA  
CATALOG



TOOL  
CATALOG



GEODATA  
CATALOG

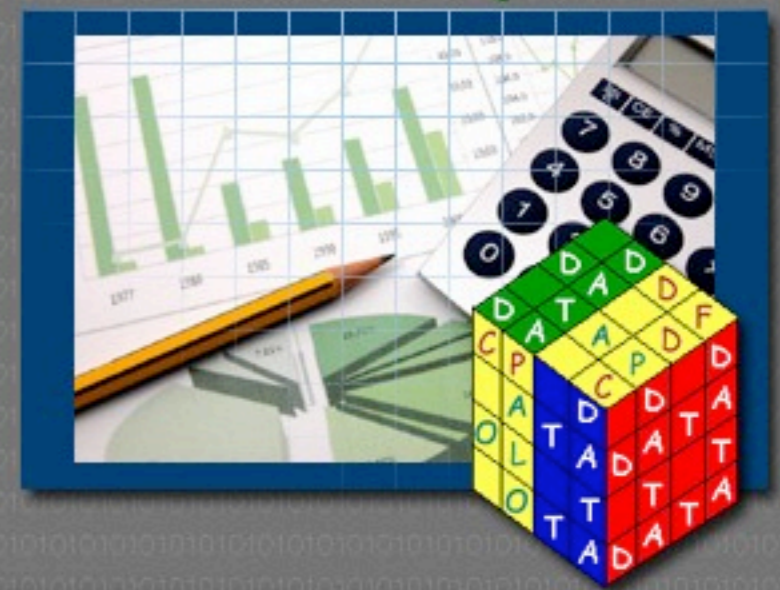
## FEATURED TOOL:

### FEDSCOPE: EASY STATS ON THE FEDERAL WORKFORCE

## FedScope

The Office of Personnel Management provides statistical information about the Federal civilian workforce. FedScope was launched in the fall of 2000. This online tool allows customers to access and analyze the most popular data elements from OPM's Central Personnel Data File (CPDF)/Enterprise Human Resources Integration (EHRI). This self-service tool provides access to 5 years worth of employment, accession, and separation data.

[VIEW THIS TOOL](#) ▶



# What Does Your Credit-Card Company Know About You?



Thomas Hannich for The New York Times

A 2002 study of how customers of Canadian Tire were using the company's credit cards found that 2,220 of 100,000 cardholders who used their credit cards in drinking places missed four payments within the next 12 months. By contrast, only 530 of the cardholders who used their credit cards at the dentist missed four payments within the next 12 months.

By **CHARLES DUHIGG**  
Published: May 12, 2009

SIGN IN TO



## EXPERT OPINION

Contact Editor: **Brian Brannon**, [bbrannon@computer.org](mailto:bbrannon@computer.org)

# The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

**E**ugene Wigner's article "The Unreasonable Effectiveness of Mathematics in the Natural Sciences"<sup>1</sup> examines why so much of physics can be neatly explained with simple mathematical formulas

behavior. So, this corpus could serve as the basis of a complete model for certain tasks—if only we knew how to extract the model from the data.

### **Learning from Text at Web Scale**

The biggest successes in natural-language-related machine learning have been statistical methods

such as *Learning from Text at Web Scale*. A complete model for

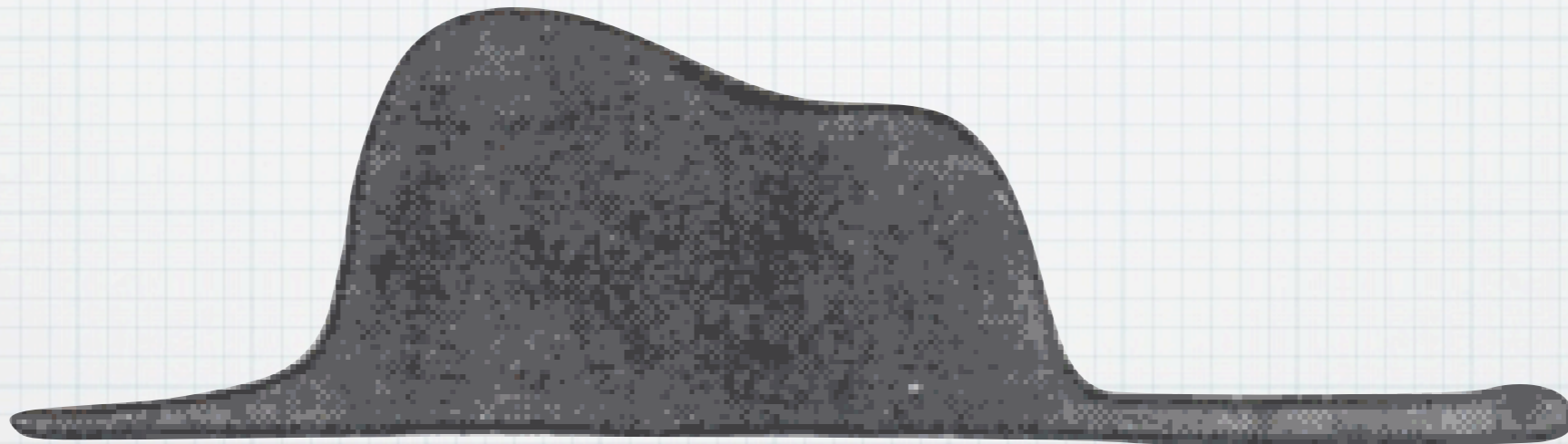
such as *Learning from Text at Web Scale*. A complete model for



“I keep saying that the sexy job  
in the next 10 years will be  
statisticians, and I’m not kidding.”

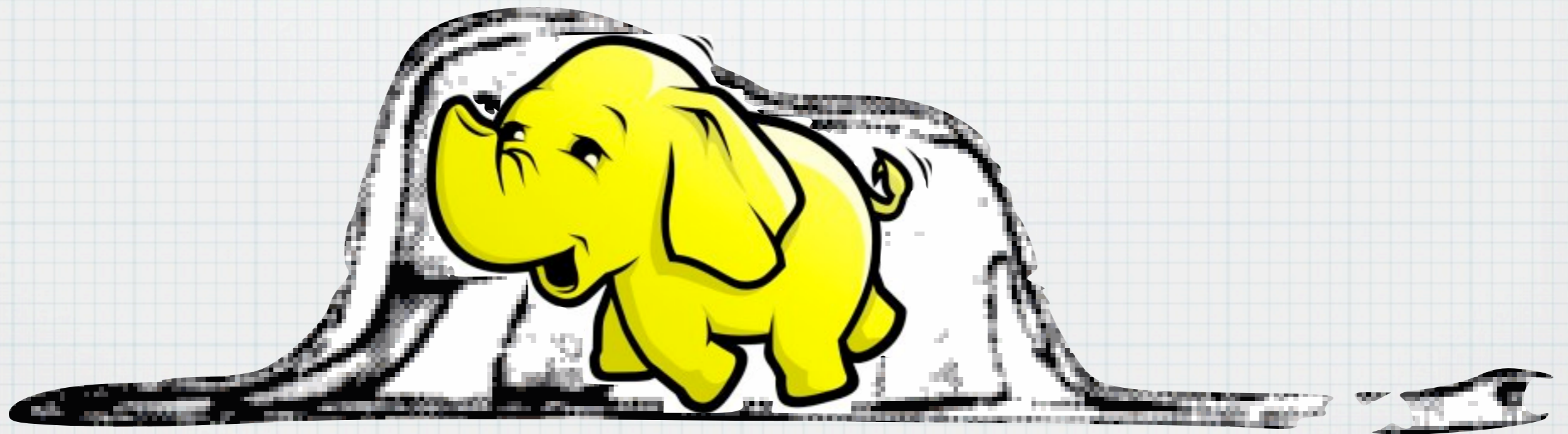
Hal Varian  
(Google’s chief economist)

# So, what's Hadoop?



*The Little Prince*, Antoine de Saint-Exupéry, Irene Testot-Ferry

Apache Hadoop is an *open-source*  
system (written in Java!) to store and  
process  
*gobs of data*  
across many commodity computers.



*The Little Prince*, Antoine de Saint-Exupéry, Irene Testot-Ferry

# Two Big Components

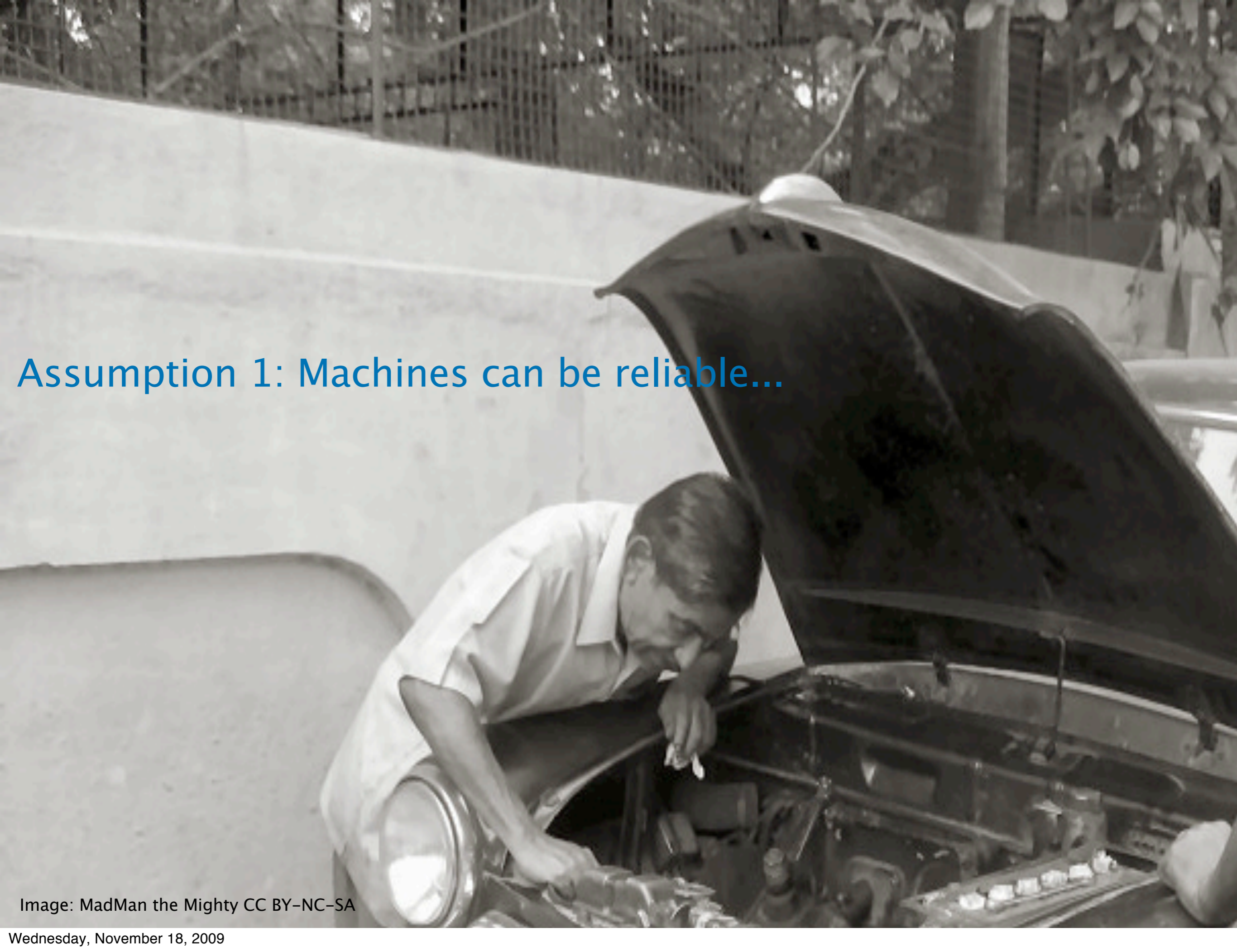
HDFS

Self-healing high-  
bandwidth  
clustered storage.

Map/Reduce

Fault-tolerant  
distributed computing.

Challenging some of  
yesteryear's  
assumptions...

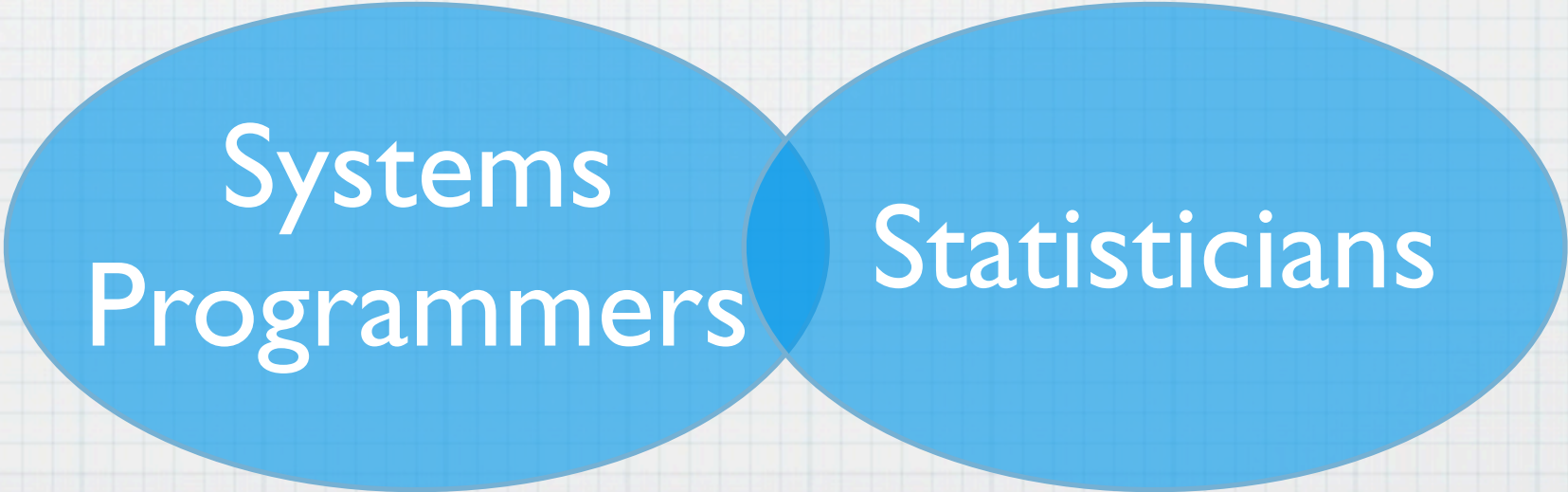


Assumption 1: Machines can be reliable...

Image: MadMan the Mighty CC BY-NC-SA

# Hadoop Goal:

Separate distributed  
system fault-tolerance  
code from application logic.



Systems  
Programmers

Statisticians



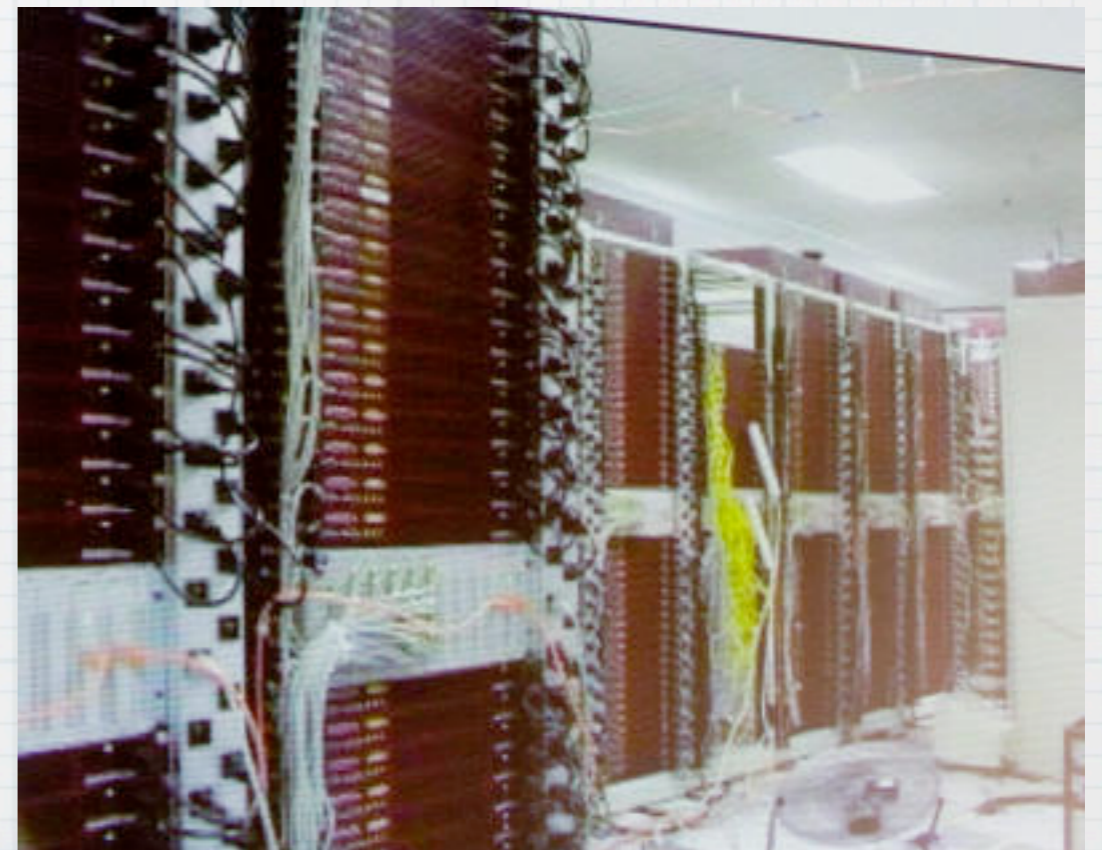
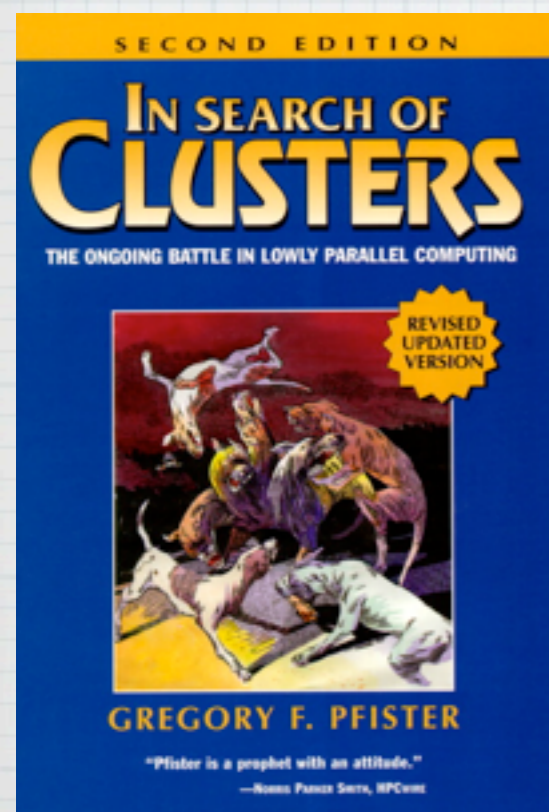
Assumption 2: Machines have identities...

Image:Laughing Squid CC BY-NC-SA



# Hadoop Goal:

Users should interact with clusters, not machines.





Assumption 3: A data set fits on one machine...

Image: Matthew J. Stinson CC-BY-NC

# Hadoop Goal:

System should scale  
linearly (or better) with  
data size.

# The M/R Programming Model

**MapReduce: Simplified Data Processing on Large Clusters**

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

*Google, Inc.*

You specify *map()*  
and *reduce()*  
functions.

The framework does  
the rest.

# map()

\*  $\text{map}: K_1, V_1 \rightarrow \text{list } K_2, V_2$

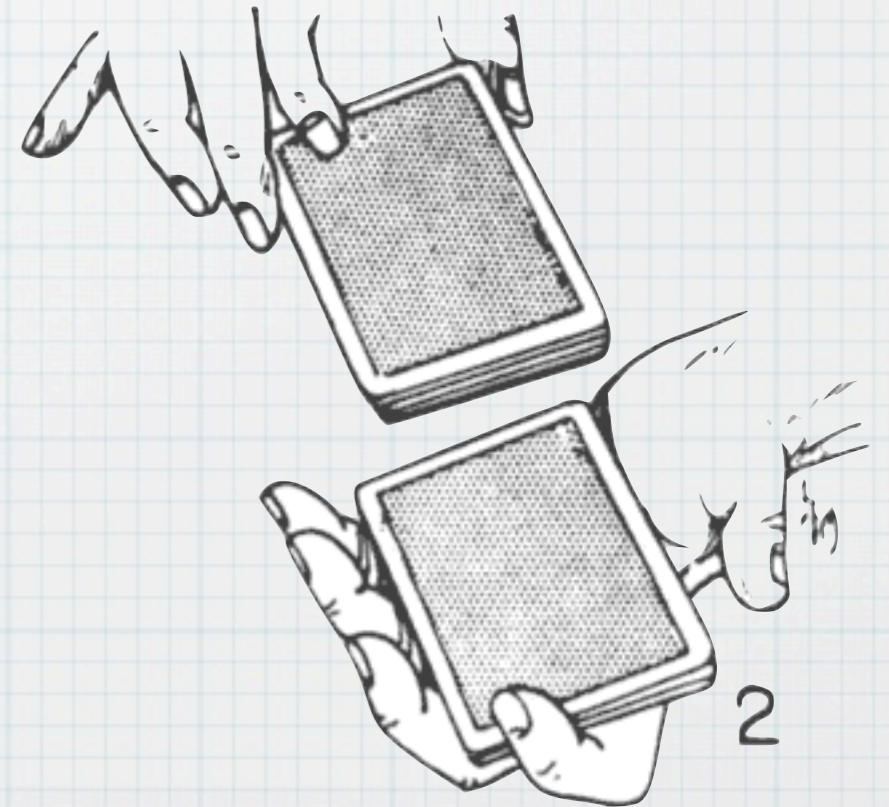
```
public class Mapper<KEYIN, VALUEIN, KEYOUT, VALUEOUT> {  
    /**  
     * Called once for each key/value pair in the input split. Most applications  
     * should override this, but the default is the identity function.  
     */  
    protected void map(KEYIN key, VALUEIN value,  
                       Context context) throws IOException,  
                       InterruptedException {  
        // context.write() can be called many times  
        // this is default "identity mapper" implementation  
        context.write((KEYOUT) key, (VALUEOUT) value);  
    }  
}
```

# (the shuffle)



\* map output is assigned to a “reducer”

\* map output is sorted by key



# reduce()

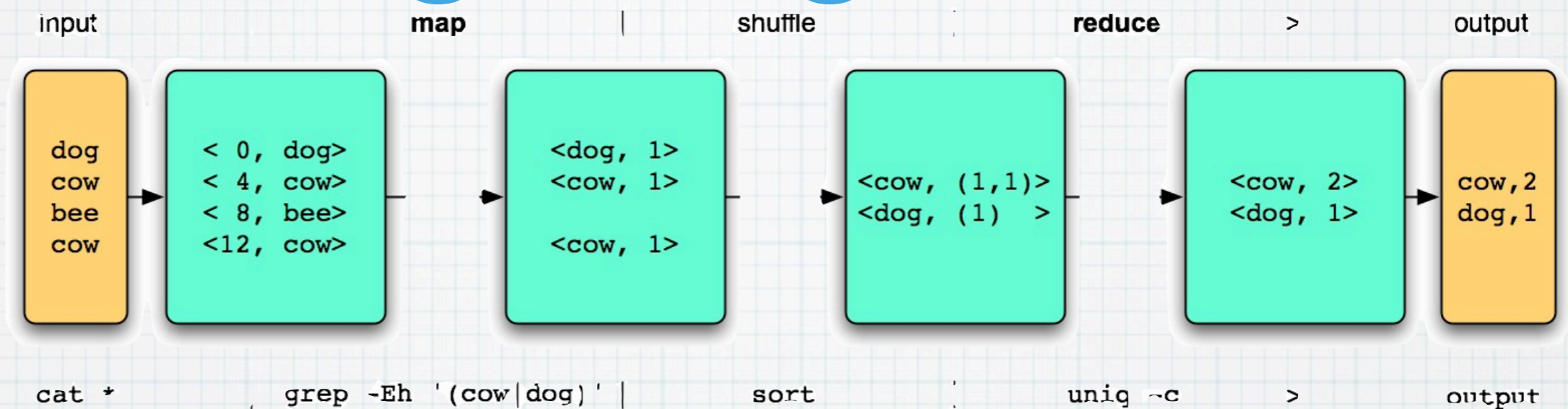
\*  $K_2, \text{iter}(V_2) \rightarrow \text{list}(K_3, V_3)$

```
public class Reducer<KEYIN, VALUEIN, KEYOUT, VALUEOUT> {  
    /**  
     * This method is called once for each key. Most applications will define  
     * their reduce class by overriding this method. The default implementation  
     * is an identity function.  
     */  
    @SuppressWarnings("unchecked")  
    protected void reduce(KEYIN key, Iterable<VALUEIN> values, Context context  
                          ) throws IOException, InterruptedException {  
        for(VALUEIN value: values) {  
            context.write((KEYOUT) key, (VALUEOUT) value);  
        }  
    }  
}
```

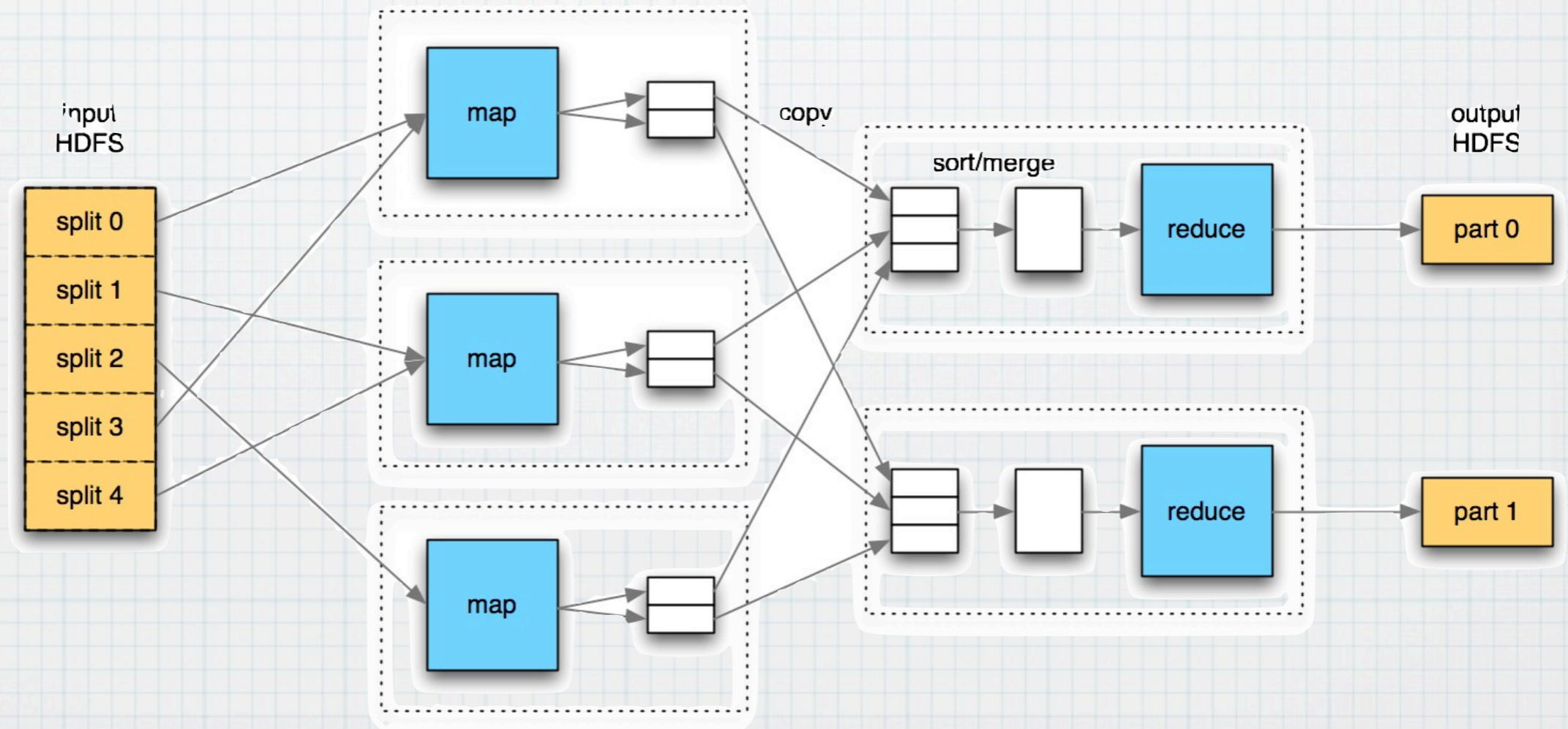


# Putting it together...

Logical



Physical



# Some samples...

- \* Build an inverted index.
- \* Summarize data grouped by a key.
- \* Build map tiles from geographic data.
- \* OCRing many images.
- \* Learning ML models. (e.g., Naive Bayes for text classification)
- \* Augment traditional BI/DW technologies (by archiving raw data).

# There's more than the Java API

## Streaming

- \* perl, python, ruby, whatever.
- \* stdin/stdout/stderr

## Pig

- \* Higher-level dataflow language for easy ad-hoc analysis.
- \* Developed at Yahoo!

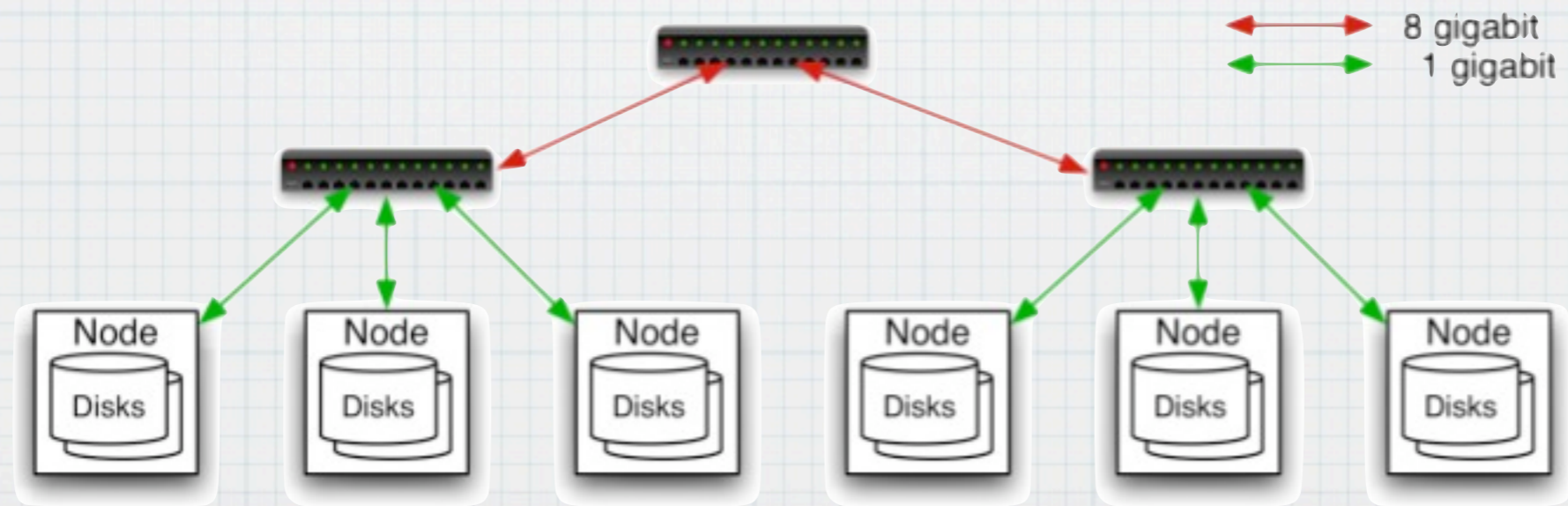
## Hive

- \* SQL interface.
- \* Great for analysts.
- \* Developed at Facebook

Friday,  
@10:10

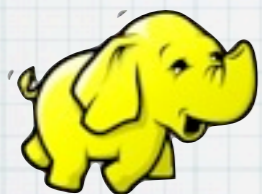
# A typical look...

- \* Commodity servers (8-core, 8-16GB RAM, 4-12 TB, 2x1 gE NIC)
- \* 2-level network architecture
- \* 20-40 nodes per rack

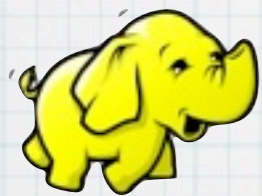


# The cast...

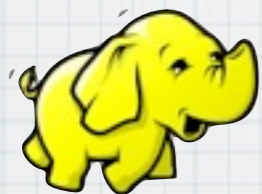
Starring...



NameNode (metadata server and database)



SecondaryNameNode (assistant to NameNode)



JobTracker (scheduler)

The Chorus...



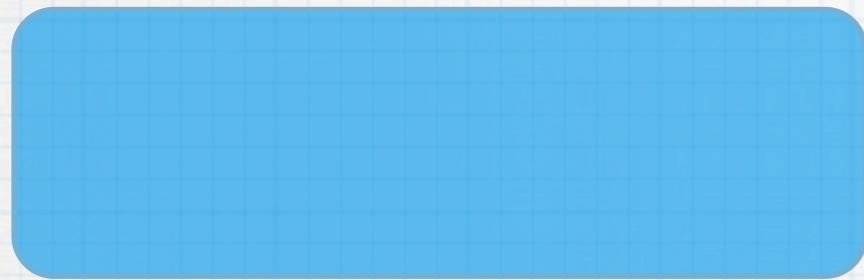
DataNodes  
(block storage)



TaskTrackers  
(task execution)

Thanks to Zak Stone for earmuff image!

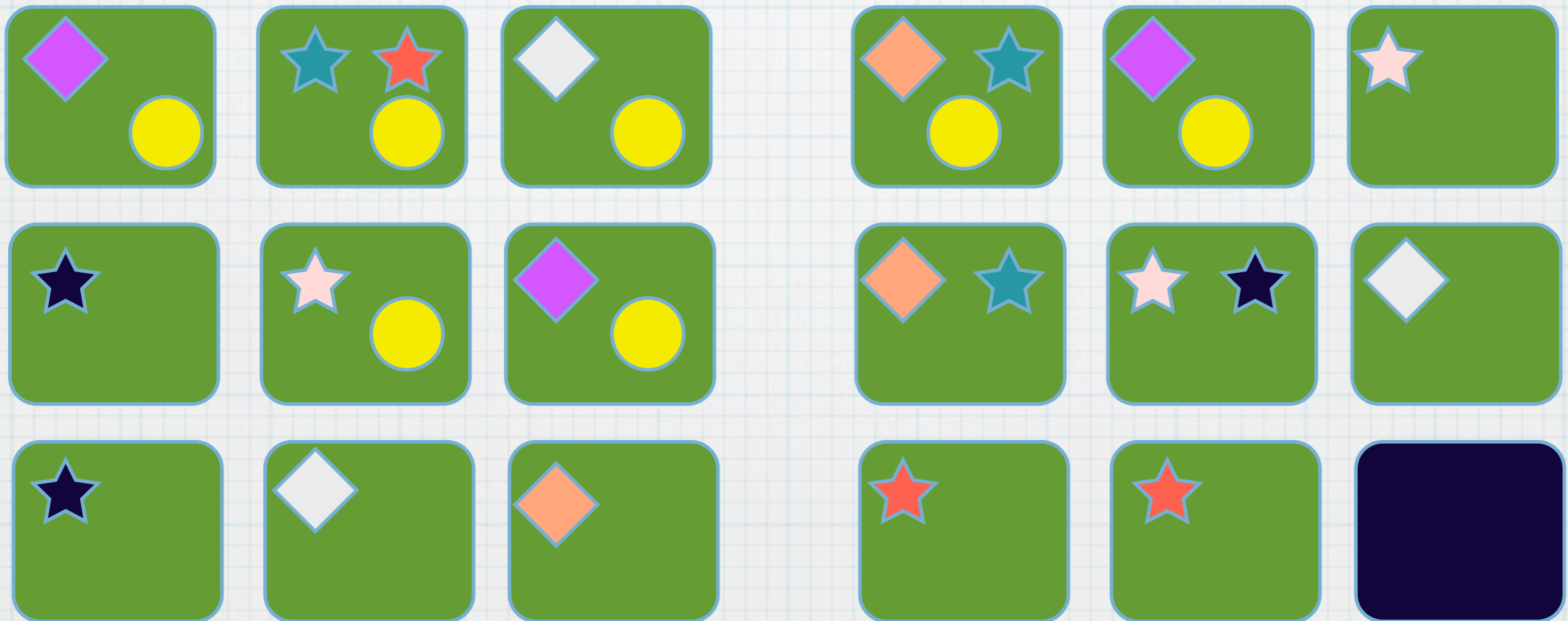
# HDFS



Namenode

- ◇ 3x64MB file, 3 rep
- ★ 4x64MB file, 3 rep
- Small file, 7 rep

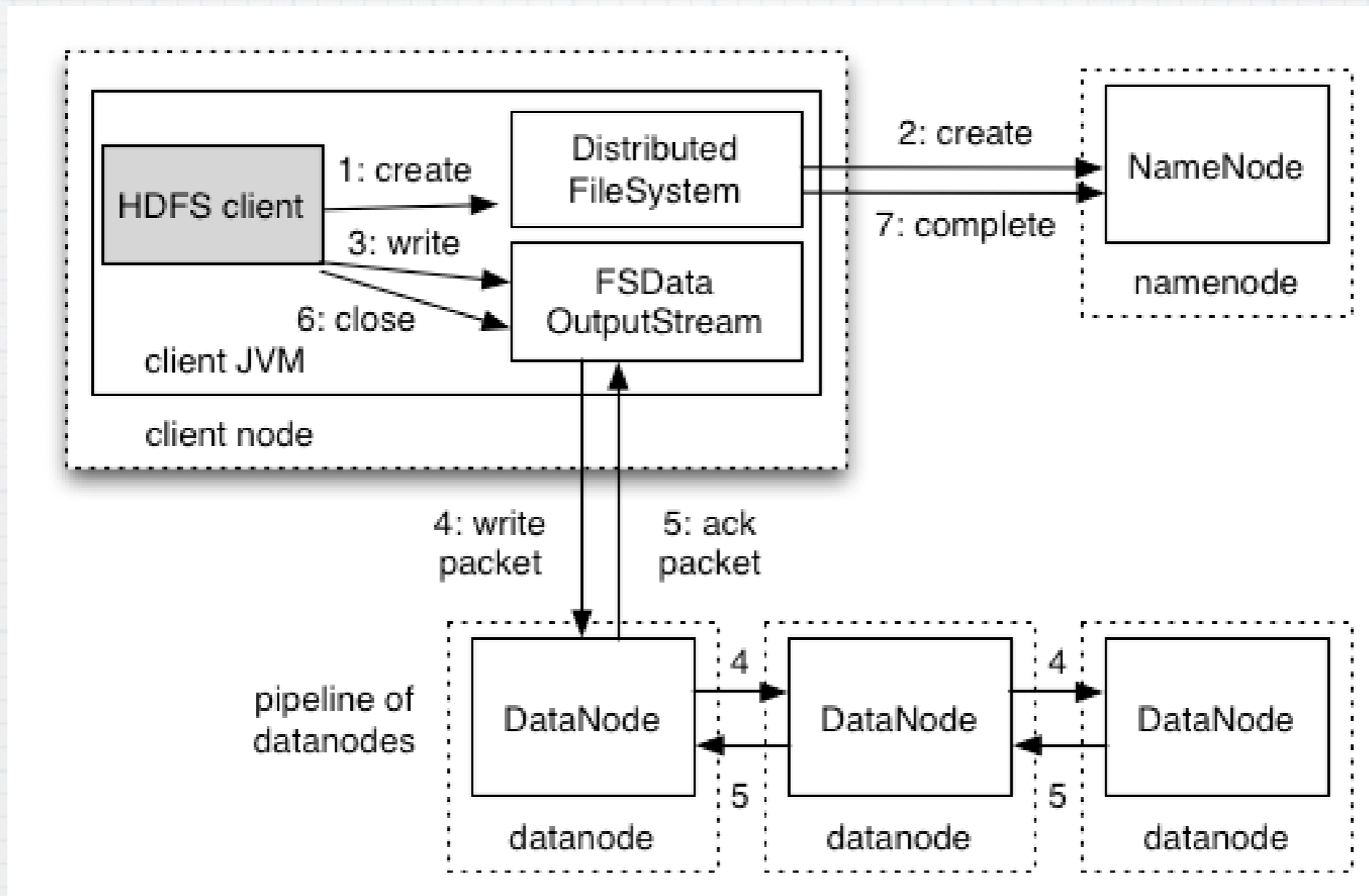
Datanodes



One Rack

A Different Rack

# HDFS Write Path



# HDFS Failures?

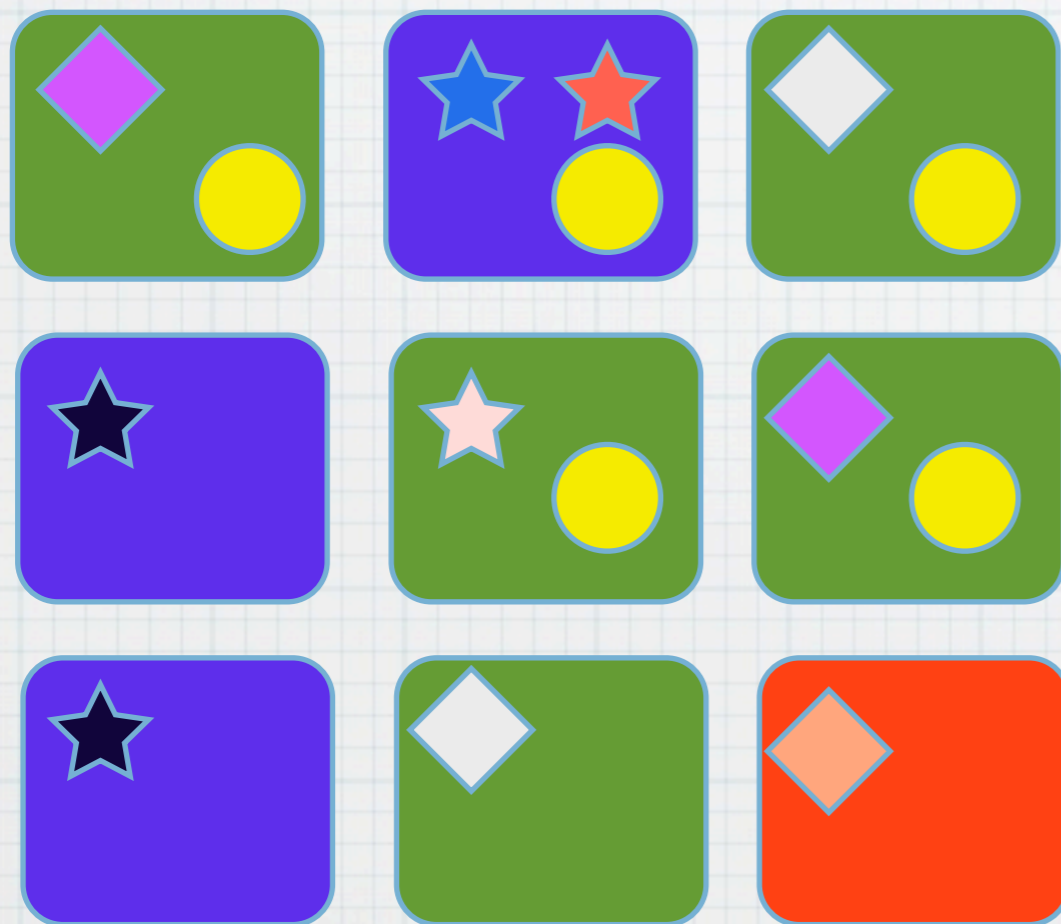
- \* Datanode crash?
  - \* Clients read another copy
  - \* Background rebalance
- \* Namenode crash?
  - \* uh-oh



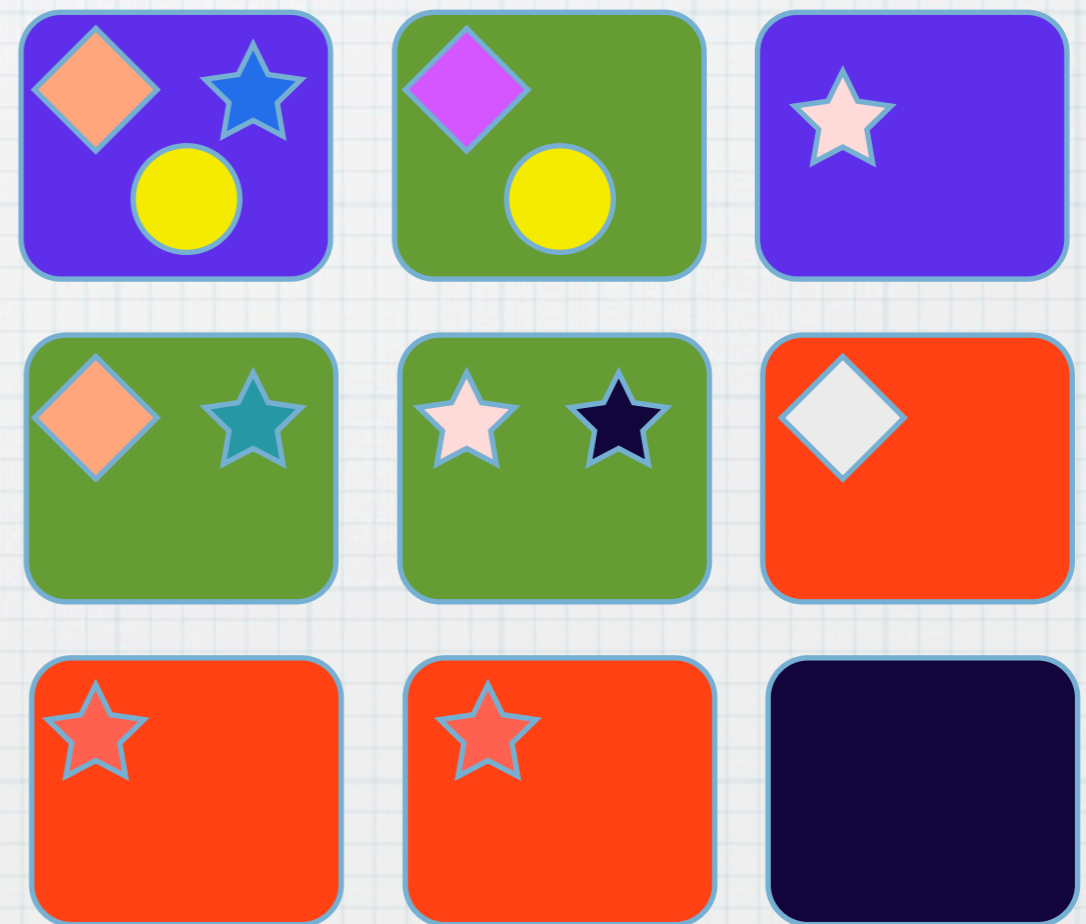


# M/R

Tasktrackers on the same machines as datanodes

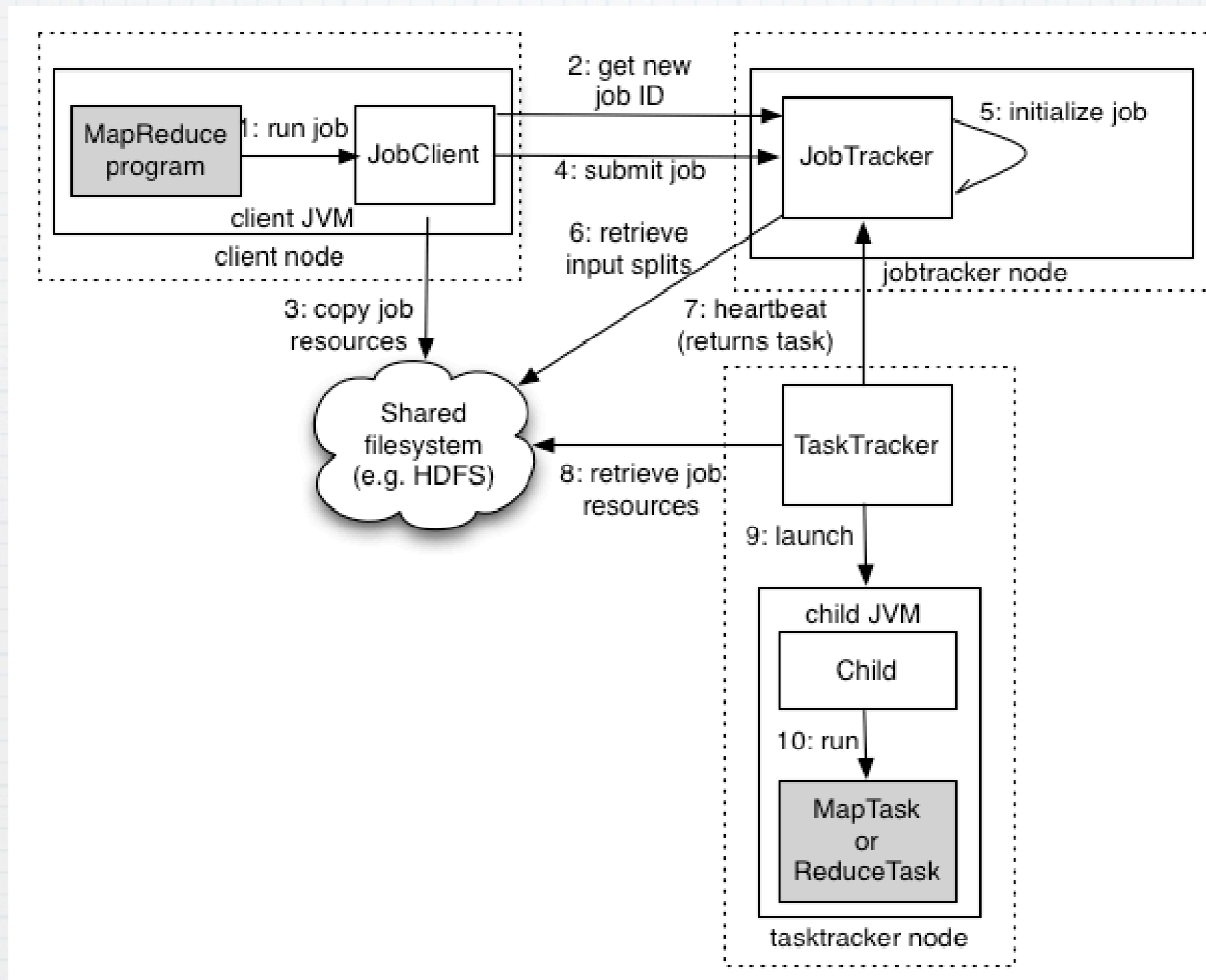


One Rack



A Different Rack

# M/R



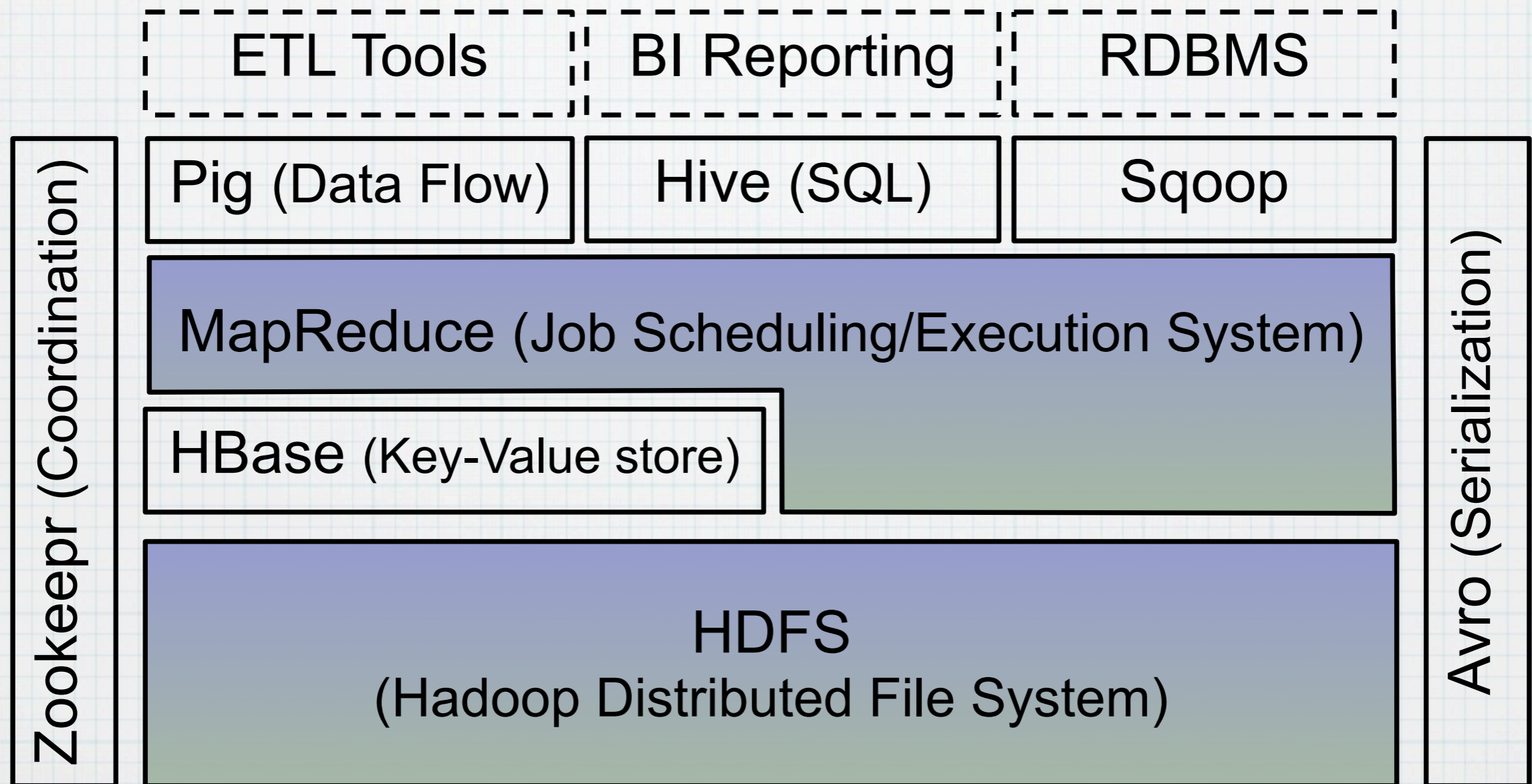
# M/R Failures

- \* Task fails
  - \* Try again?
  - \* Try again somewhere else?
  - \* Report failure
- \* Retries possible because of *idempotence*

# Hadoop in the Wild

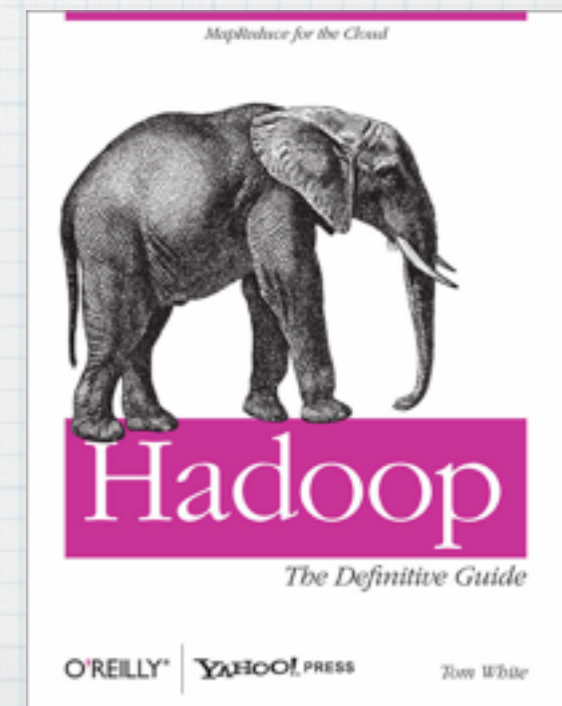
- \* Yahoo! Hadoop Clusters: > 82PB, >25k machines (Eric I 4, HadoopWorld NYC '09)
- \* Google: 40 GB/s GFS read/write load (Jeff Dean, LADIS '09) [ $\sim$ 3,500 TB/day]
- \* Facebook: 4TB new data per day; DW: 4800 cores, 5.5 PB (Dhruba Borthakur, HadoopWorld)

# The Hadoop Ecosystem



# Ok, fine, what next?

- \* Get Hadoop!
- \* <http://hadoop.apache.org/>
- \* Cloudera Distribution for Hadoop
- \* Try it out! (Locally, or on EC2)



# Just one slide...



# cloudera

- \* Software: Cloudera Distribution for Hadoop, Cloudera Desktop, more...
- \* Training and certification...
- \* Free on-line training materials (including video)
- \* Support & Professional Services
- \* @cloudera, blog, etc.

# Questions?

[philip@cloudera.com](mailto:philip@cloudera.com)

