Lecture 17:

Spark

(leveraging bulk-granularity program structure)

Parallel Computer Architecture and Programming

CMU / 清华大学, Summer 2017

Review: which program performs better?

Program 1

```
void add(int n, float* A, float* B, float* C) {
    for (int i=0; i<n; i++)
       C[i] = A[i] + B[i];
}

void mul(int n, float* A, float* B, float* C) {
    for (int i=0; i<n; i++)
       C[i] = A[i] * B[i];
}


float* A, *B, *C, *D, *E, *tmp1, *tmp2;

// assume arrays are allocated here

// compute E = D + ((A + B) * C)
add(n, A, B, tmp1);
mul(n, tmp1, C, tmp2);
add(n, tmp2, D, E);
```

Two loads, one store per math op (arithmetic intensity = 1/3)

Two loads, one store per math op (arithmetic intensity = 1/3)

Overall arithmetic intensity = 1/3

Program 2

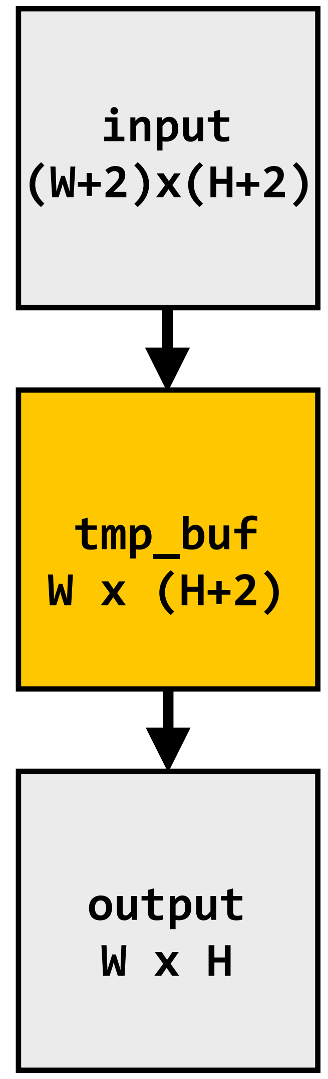
```
void fused(int n, float* A, float* B, float* C, float* D, float* E) {
    for (int i=0; i<n; i++)
       E[i] = D[i] + (A[i] + B[i]) * C[i];
}

// compute E = D + (A + B) * C
fused(n, A, B, C, D, E);
```

Four loads, one store per 3 math ops (arithmetic intensity = 3/5)

The transformation of the code in program 1 to the code in program 2 is called "loop fusion"

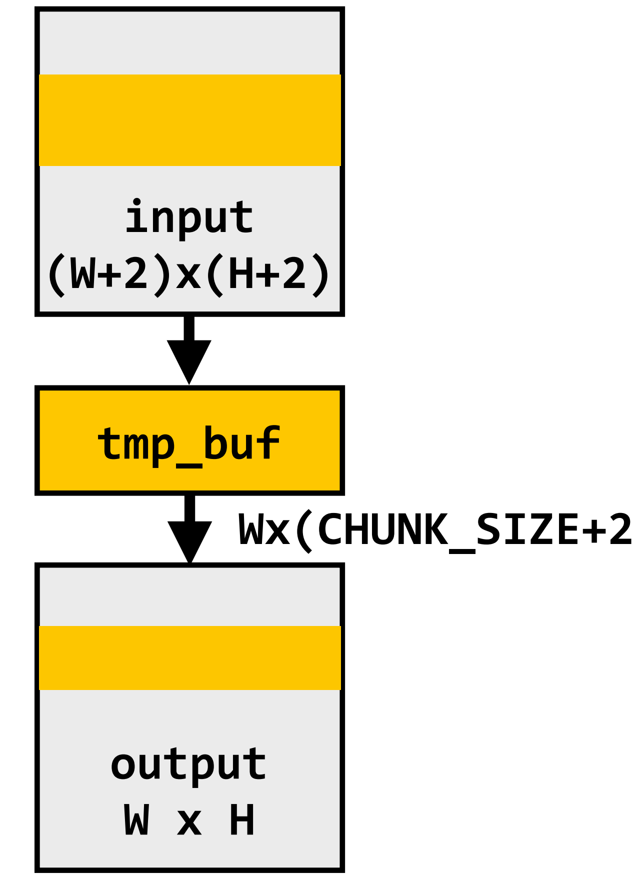
Review: why did we perform this transform?

Program 1




```
int WIDTH = 1024;
int HEIGHT = 1024;
float input[(WIDTH+2) * (HEIGHT+2)];
float tmp_buf[WIDTH * (HEIGHT+2)];
float output[WIDTH * HEIGHT];

float weights[] = {1.0/3, 1.0/3, 1.0/3};

// blur image horizontally
for (int j=0; j<(HEIGHT+2); j++)
  for (int i=0; i<WIDTH; i++) {
    float tmp = 0.f;
    for (int ii=0; ii<3; ii++)
      tmp += input[j*(WIDTH+2) + i+ii] * weights[ii];
    tmp_buf[j*WIDTH + i] = tmp;
  }

// blur tmp_buf vertically
for (int j=0; j<HEIGHT; j++) {
  for (int i=0; i<WIDTH; i++) {
    float tmp = 0.f;
    for (int jj=0; jj<3; jj++)
      tmp += tmp_buf[(j+jj)*WIDTH + i] * weights[jj];
    output[j*WIDTH + i] = tmp;
  }
}
```

Program 2




```
int WIDTH = 1024;
int HEIGHT = 1024;
float input[(WIDTH+2) * (HEIGHT+2)];
float tmp_buf[WIDTH * (CHUNK_SIZE+2)];
float output[WIDTH * HEIGHT];

float weights[] = {1.0/3, 1.0/3, 1.0/3};

for (int j=0; j<HEIGHT; j+CHUNK_SIZE) {

  // blur region of image horizontally
  for (int j2=0; j2<CHUNK_SIZE+2; j2++)
    for (int i=0; i<WIDTH; i++) {
      float tmp = 0.f;
      for (int ii=0; ii<3; ii++)
        tmp += input[(j+j2)*(WIDTH+2) + i+ii] * weights[ii];
      tmp_buf[j2*WIDTH + i] = tmp;

  // blur tmp_buf vertically
  for (int j2=0; j2<CHUNK_SIZE; j2++)
    for (int i=0; i<WIDTH; i++) {
      float tmp = 0.f;
      for (int jj=0; jj<3; jj++)
        tmp += tmp_buf[(j2+jj)*WIDTH + i] * weights[jj];
      output[(j+j2)*WIDTH + i] = tmp;
    }
}
```

Both of the previous examples involved globally changing the order of computation to improve producer-consumer locality

(improve arithmetic intensity of program)

A log of page views on the class web site

```
537.36 (KHTML, like Gecko) Chrome/59.0.3071.115 Safari/537.36"
207.46.13.25 - - [19/Jul/2017:18:07:08 -0400] "GET /fall2016content/lectures/12_snoopimpl/12_snoopimpl_slides.pdf HTTP/1.1" 200 2256981 "-" "Mozilla/5.0 (iPhone; CPU iPhone OS 7_0 like Mac OS X)
AppleWebKit/537.51.1 (KHTML, like Gecko) Version/7.0 Mobile/11A465 Safari/9537.53 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)"
166.111.131.62 - - [19/Jul/2017:18:07:44 -0400] "GET /tsinghua2017/keep_alive HTTP/1.1" 200 914 "http://15418.courses.cs.cmu.edu/tsinghua2017/assignments" "Mozilla/5.0 (X11; Linux x86_64; rv:52.0) Ge
20100101 Firefox/52.0"
166.111.131.62 - - [19/Jul/2017:18:07:54 -0400] "GET /tsinghua2017/keep_alive HTTP/1.1" 200 413 "http://15418.courses.cs.cmu.edu/tsinghua2017/article/3" "Mozilla/5.0 (X11; Linux x86_64; rv:52.0) Geck
20100101 Firefox/52.0"
207.46.13.25 - - [19/Jul/2017:18:11:53 -0400] "GET /fall2016content/lectures/05_progperf1/05_progperf1_slides.pdf HTTP/1.1" 200 1654279 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com
bingbot.htm)"
157.55.39.45 - - [19/Jul/2017:18:12:34 -0400] "GET /spring2015/lecture/progmodels/slide_012 HTTP/1.1" 200 4737 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)"
165.230.224.216 - - [19/Jul/2017:18:12:38 -0400] "GET /spring2017/keep_alive HTTP/1.1" 200 921 "http://15418.courses.cs.cmu.edu/spring2017/home" "Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:54.0) Geck
20100101 Firefox/54.0"
172.31.51.99 - - [19/Jul/2017:18:13:22 -0400] "GET /spring2017/keep_alive HTTP/1.1" 200 964 "http://15418.courses.cs.cmu.edu/spring2017/lecture/progperf1" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_1
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/59.0.3071.115 Safari/537.36"
180.169.142.74 - - [19/Jul/2017:18:14:09 -0400] "GET /spring2017/keep_alive HTTP/1.1" 200 952 "http://15418.courses.cs.cmu.edu/spring2017/" "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML,
like Gecko) Chrome/59.0.3071.104 Safari/537.36"
202.79.36.139 - - [19/Jul/2017:18:15:08 -0400] "GET /wp-login.php HTTP/1.1" 404 540 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64; rv:40.0) Gecko/20100101 Firefox/40.1"
202.79.36.139 - - [19/Jul/2017:18:15:09 -0400] "GET / HTTP/1.1" 302 605 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64; rv:40.0) Gecko/20100101 Firefox/40.1"
202.79.36.139 - - [19/Jul/2017:18:15:09 -0400] "GET /spring2017 HTTP/1.1" 301 643 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64; rv:40.0) Gecko/20100101 Firefox/40.1"
202.79.36.139 - - [19/Jul/2017:18:15:10 -0400] "GET /spring2017/ HTTP/1.1" 200 19948 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64; rv:40.0) Gecko/20100101 Firefox/40.1"
75.110.30.93 - - [19/Jul/2017:18:15:29 -0400] "GET /spring2016/keep_alive HTTP/1.1" 200 965 "http://15418.courses.cs.cmu.edu/spring2016/home" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12_5) AppleWeb
537.36 (KHTML, like Gecko) Chrome/59.0.3071.115 Safari/537.36"
75.110.30.93 - - [19/Jul/2017:18:15:31 -0400] "GET /spring2015/keep_alive HTTP/1.1" 200 964 "http://15418.courses.cs.cmu.edu/spring2015/reading" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12_5)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/59.0.3071.115 Safari/537.36"
166.111.131.62 - - [19/Jul/2017:18:15:31 -0400] "GET /tsinghua2017/keep_alive HTTP/1.1" 200 924 "http://15418.courses.cs.cmu.edu/tsinghua2017/assignments" "Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:
54.0) Gecko/20100101 Firefox/54.0"
128.237.28.16 - - [19/Jul/2017:18:16:06 -0400] "GET /spring2014content/lectures/04_progbasics/thumbs/slide_013.png HTTP/1.1" 304 189 "-" "Mozilla/4.0 (compatible;)"
24.176.186.198 - - [19/Jul/2017:18:16:10 -0400] "GET /spring2015/keep_alive HTTP/1.1" 200 964 "http://15418.courses.cs.cmu.edu/spring2015/" "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.
(KHTML, like Gecko) Chrome/59.0.3071.115 Safari/537.36"
89.191.14.105 - - [19/Jul/2017:18:16:30 -0400] "GET /spring2017/keep_alive HTTP/1.1" 200 962 "http://15418.courses.cs.cmu.edu/spring2017/" "Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/59.0.3071.115 Safari/537.36"
141.8.143.158 - - [19/Jul/2017:18:17:25 -0400] "GET /spring2017/lecture/dsl/slide_050 HTTP/1.1" 200 3413 "-" "Mozilla/5.0 (compatible; YandexBot/3.0; +http://yandex.com/bots)"
141.8.143.158 - - [19/Jul/2017:18:17:32 -0400] "GET /spring2017/lecture/perfeval/slide_016 HTTP/1.1" 200 3621 "-" "Mozilla/5.0 (compatible; YandexBot/3.0; +http://yandex.com/bots)"
141.8.143.220 - - [19/Jul/2017:18:17:59 -0400] "GET /spring2017/lecture/synchronization/slide_036 HTTP/1.1" 200 3390 "-" "Mozilla/5.0 (compatible; YandexBot/3.0; +http://yandex.com/bots)"
165.204.55.250 - - [19/Jul/2017:18:18:24 -0400] "GET /spring2015/keep_alive HTTP/1.1" 200 969 "http://15418.courses.cs.cmu.edu/spring2015/competition" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit
537.36 (KHTML, like Gecko) Chrome/58.0.3029.110 Safari/537.36 OPR/45.0.2552.898"
141.8.143.220 - - [19/Jul/2017:18:18:26 -0400] "GET /spring2017/lecture/lockfree/slide_031 HTTP/1.1" 200 5224 "-" "Mozilla/5.0 (compatible; YandexBot/3.0; +http://yandex.com/bots)"
64.79.115.181 - - [19/Jul/2017:18:18:30 -0400] "GET /spring2015/keep_alive HTTP/1.1" 200 966 "http://15418.courses.cs.cmu.edu/spring2015/" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_6) AppleWebKit
537.36 (KHTML, like Gecko) Chrome/58.0.3029.96 Safari/537.36"
166.111.131.62 - - [19/Jul/2017:18:18:57 -0400] "GET /tsinghua2017/keep_alive HTTP/1.1" 200 914 "http://15418.courses.cs.cmu.edu/tsinghua2017/lecture/scheduling" "Mozilla/5.0 (X11; Linux x86_64; rv:5
Gecko/20100101 Firefox/52.0"
141.8.143.220 - - [19/Jul/2017:18:19:14 -0400] "GET /spring2017/lecture/graphdsl/slide_036 HTTP/1.1" 200 3241 "-" "Mozilla/5.0 (compatible; YandexBot/3.0; +http://yandex.com/bots)"
141.8.143.220 - - [19/Jul/2017:18:19:16 -0400] "GET /spring2013/lecture/synchronization HTTP/1.1" 200 21958 "-" "Mozilla/5.0 (compatible; YandexBot/3.0; +http://yandex.com/bots)"
94.153.230.50 - - [19/Jul/2017:18:20:34 -0400] "GET /spring2016/keep_alive HTTP/1.1" 200 876 "http://15418.courses.cs.cmu.edu/spring2016/" "Mozilla/5.0 (Windows NT 6.1; WOW64; rv:53.0) Gecko/20100101
Firefox/53.0"
157.55.39.45 - - [19/Jul/2017:18:20:41 -0400] "GET /spring2016/lecture/synchronization/slide_034 HTTP/1.1" 200 4833 "-" "Mozilla/5.0 (iPhone; CPU iPhone OS 7_0 like Mac OS X) AppleWebKit/537.51.1 (KH
like Gecko) Version/7.0 Mobile/11A465 Safari/9537.53 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)"
165.204.55.250 - - [19/Jul/2017:18:22:51 -0400] "GET /spring2016/keep_alive HTTP/1.1" 200 969 "http://15418.courses.cs.cmu.edu/spring2016/exercises" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/
537.36 (KHTML, like Gecko) Chrome/58.0.3029.110 Safari/537.36 OPR/45.0.2552.898"
141.8.143.220 - - [19/Jul/2017:18:23:40 -0400] "GET /spring2014/lecture/progbasics/slide_001 HTTP/1.1" 200 3703 "-" "Mozilla/5.0 (compatible; YandexBot/3.0; +http://yandex.com/bots)"
157.55.39.45 - - [19/Jul/2017:18:23:56 -0400] "GET /spring2017/lecture/dnn/slide_025 HTTP/1.1" 200 3145 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)"
166.111.131.62 - - [19/Jul/2017:18:24:01 -0400] "GET /tsinghua2017/keep_alive HTTP/1.1" 200 924 "http://15418.courses.cs.cmu.edu/tsinghua2017/article/4" "Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:54
Gecko/20100101 Firefox/54.0"
141.8.143.220 - - [19/Jul/2017:18:24:17 -0400] "GET /spring2017/lecture/lockfree/slide_018 HTTP/1.1" 200 4102 "-" "Mozilla/5.0 (compatible; YandexBot/3.0; +http://yandex.com/bots)"
207.46.13.25 - - [19/Jul/2017:18:24:42 -0400] "GET /spring2017/article/13 HTTP/1.1" 200 9243 "-" "Mozilla/5.0 (iPhone; CPU iPhone OS 7_0 like Mac OS X) AppleWebKit/537.51.1 (KHTML, like Gecko) Versio
7.0 Mobile/11A465 Safari/9537.53 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)"
```

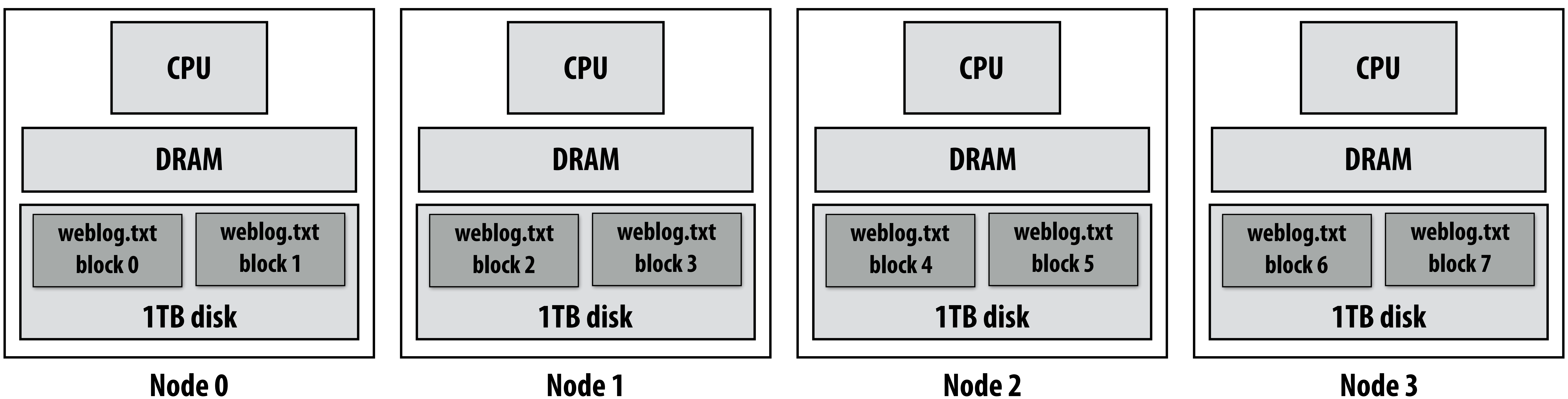
Parallel programming is very popular…

The log of page views gets quite large…

Assume weblog.txt is a large file, stored in a distributed file system, like HDFS

Below: cluster of four nodes, each node with a 1 TB disk

Contents of weblog.txt are distributed evenly in blocks across the cluster



Imagine that the agents for the bands want to know more about the fans reading the parallel programming class web site…

For example:
"What type of mobile phone are all the fans using?"

A simple programming model

```
// called once per line in input file by runtime
// input:  string (one line from input file)
// output: appends (user_agent, 1) entry to results list
void mapper(string line, multimap<string,string>& results) {
   string user_agent = parse_requester_user_agent(line);
   if (is_mobile_device(user_agent))
     results.add(user_agent, 1);
}

// called once per unique key (user_agent) in results
// values is a list of values associated with the given key
void reducer(string key, list<string> values, int& result) {
    int sum = 0;
    for (v in values)
       sum += v;
    result = sum;
}


// iterator over lines of text file
LineByLineReader input(“hdfs://weblog.txt”);

// stores output
Writer output(“hdfs://…”);

// do work
runMapReduceJob(mapper, reducer, input, output);
```

(The code above computes the count of page views by each type of mobile phone.)

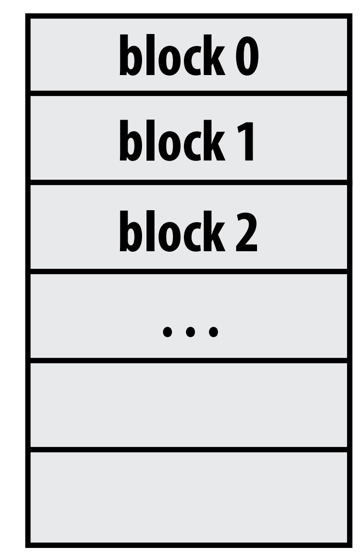
Let's design an implementation of runMapReduceJob

Step 1: running the mapper function

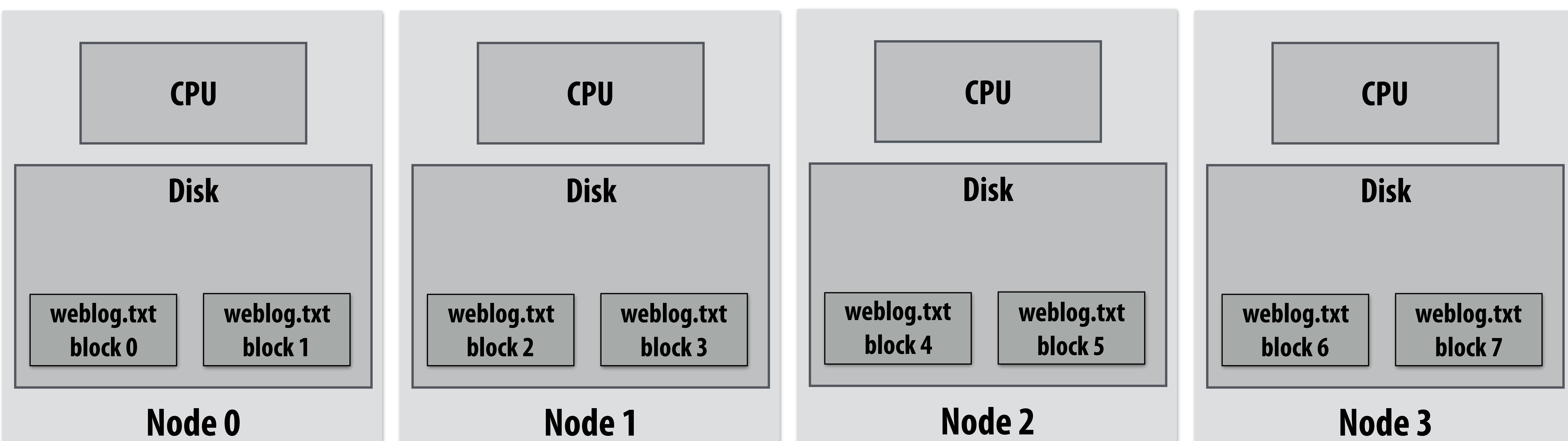
```
// called once per line in file
void mapper(string line, multimap<string,string>& results) {
   string user_agent = parse_requester_user_agent(line);
   if (is_mobile_device(user_agent))
     results.add(user_agent, 1);
}

// called once per unique key in results
void reducer(string key, list<string> values, int& result) {
    int sum = 0;
    for (v in values)
       sum += v;
    result = sum;
}

LineByLineReader input("hdfs://weblog.txt");
Writer output("hdfs://…");
runMapReduceJob(mapper, reducer, input, output);
```

Step 1: run mapper function on all lines of file
Question: How to assign work to nodes?

Idea 1: use work queue for list of input blocks to process
Dynamic assignment: free node takes next available block



Idea 2: data distribution based assignment: Each node processes lines in blocks of input file that are stored locally.



Steps 2 and 3: gathering data, running the reducer

```
// called once per line in file
void mapper(string line, map<string,string> results) {
   string user_agent = parse_requester_user_agent(line);
   if (is_mobile_device(user_agent))
     results.add(user_agent, 1);
}

// called once per unique key in results
void reducer(string key, list<string> values, int& result) {
    int sum = 0;
    for (v in values)
       sum += v;
    result = sum;
}

LineByLineReader input("hdfs://weblog.txt");
Writer output("hdfs://…");
runMapReduceJob(mapper, reducer, input, output);
```

Step 2: Prepare intermediate data for reducer

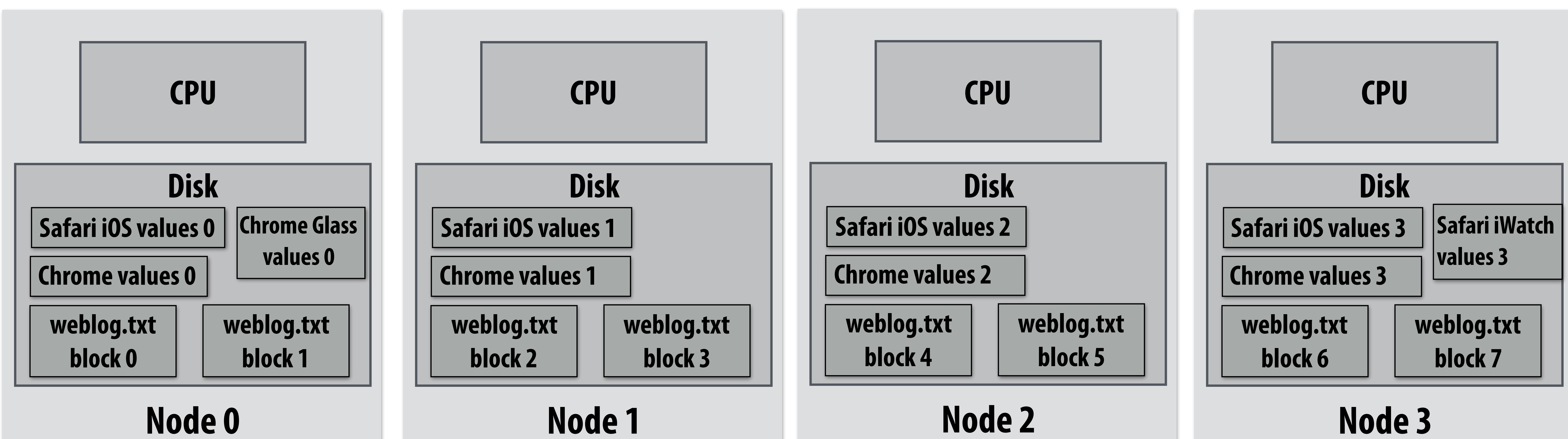
Step 3: Run reducer function on all keys

Question: how to assign reducer tasks?

Question: how to get all data for one key to the correct worker node?

Keys to reduce: (generated by mapper):

Safari iOS
Chrome
Safari iWatch
Chrome Glass



Steps 2 and 3: gathering data, running the reducer

```
// gather all input data for key, then execute reducer
// to produce final result
void runReducer(string key, reducer, result) {
   list<string> inputs;
   for (n in nodes) {
        filename = get_filename(key, n);
        read lines of filename, append into inputs;
   }
   reducer(key, inputs, result);
}
```

Step 2: Prepare intermediate data for reducer.

Step 3: Run reducer function on all keys.

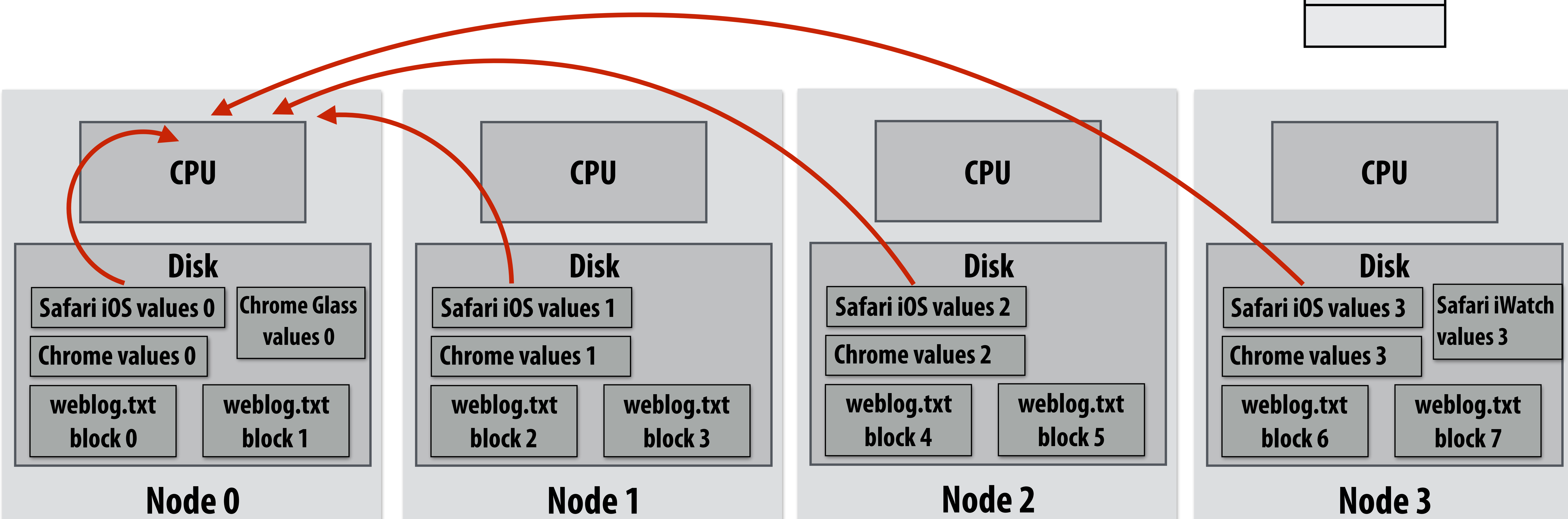
Question: how to assign reducer tasks?

Question: how to get all data for key onto the correct worker node?

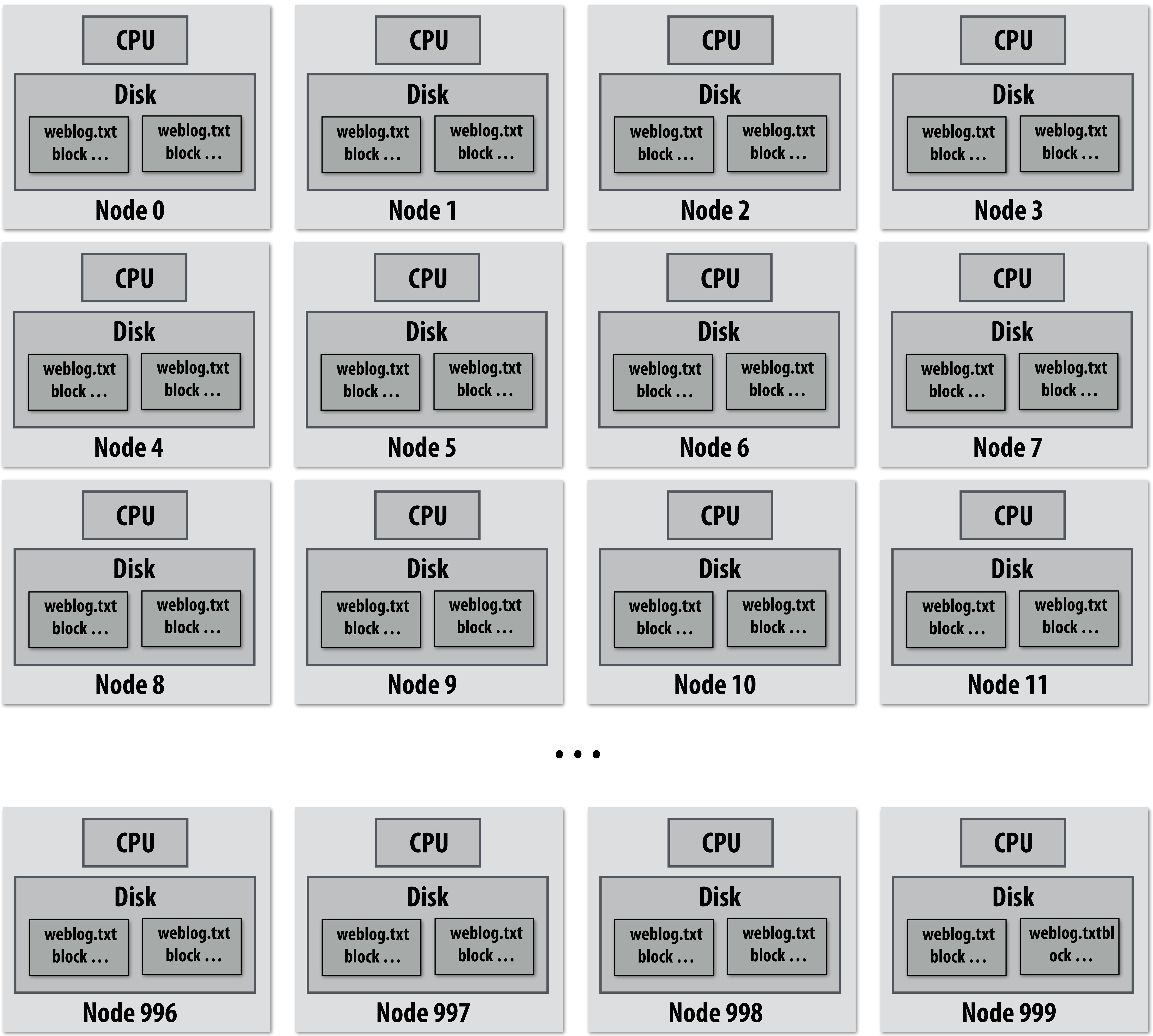
Example:
Assign Safari iOS to Node 0

Keys to reduce:
(generated by mapper):

Safari iOS
Chrome
Safari iWatch
Chrome Glass



Additional implementation challenges at scale



Nodes may fail during program execution

Some nodes may run slower than others
(due to different amounts of work, heterogeneity in the cluster, etc..)

Job scheduler responsibilities

- **Exploit data locality: "move computation to the data"**
  - Run mapper jobs on nodes that contain input files
  - Run reducer jobs on nodes that already have most of data for a certain key

- **Handling node failures**
  - Scheduler detects job failures and reruns job on new machines
    - This is possible since inputs reside in persistent storage (distributed file system)
  - Scheduler duplicates jobs on multiple machines (reduce overall processing latency incurred by node failures)

- **Handling slow machines**
  - Scheduler may even duplicate jobs on multiple machines in case one runs slow

runMapReduceJob problems?

- **Permits only a very simple program structure**
  - Programs must be structured as: map, followed by reduce by key
  - See DryadLINQ for generalization to DAGs

- **Iterative algorithms must load from disk each iteration**
  - Recall last lecture on graph processing:

```
void pagerank_mapper(graphnode n, map<string,string> results) {
   float val = compute update value for n
   for (dst in outgoing links from n)
     results.add(dst.node, val);
}

void pagerank_reducer(graphnode n, list<float> values, float& result) {
    float sum = 0.0;
    for (v in values)
       sum += v;
    result = sum;
}

for (i = 0 to NUM_ITERATIONS) {
   input = load graph from last iteration
   output = file for this iteration output
   runMapReduceJob(pagerank_mapper, pagerank_reducer, result[i-1], result[i]);
}
```

in-memory, fault-tolerant distributed computing

http://spark.apache.org/

[Zaharia et al. NSDI 2012]

Goals

- Programming model for cluster-scale computations where there is significant reuse of intermediate datasets
  - Iterative machine learning and graph algorithms
  - Interactive data mining: load large dataset into aggregate memory of cluster and then perform multiple ad-hoc queries

- Don't want incur inefficiency of writing intermediates to persistent distributed file system (want to keep it in memory)
  - Challenge: efficiently implementing fault tolerance for large-scale distributed in-memory computations.

Fault tolerance for in-memory calculations

- **Replicate all computations**
  - Expensive solution: decreases peak throughput
- **Checkpoint and rollback**
  - Periodically save state of program to persistent storage
  - Restart from last checkpoint on node failure
- **Maintain log of updates (commands and data)**
  - High overhead for maintaining logs

Recall map-reduce solutions:

- Checkpoints after each map/reduce step by writing results to file system
- Scheduler's list of outstanding (but not yet complete) jobs is a log
- Functional structure of programs allows for restart at granularity of a single mapper or reducer invocation (don't have to restart entire program)

Resilient distributed dataset (RDD)

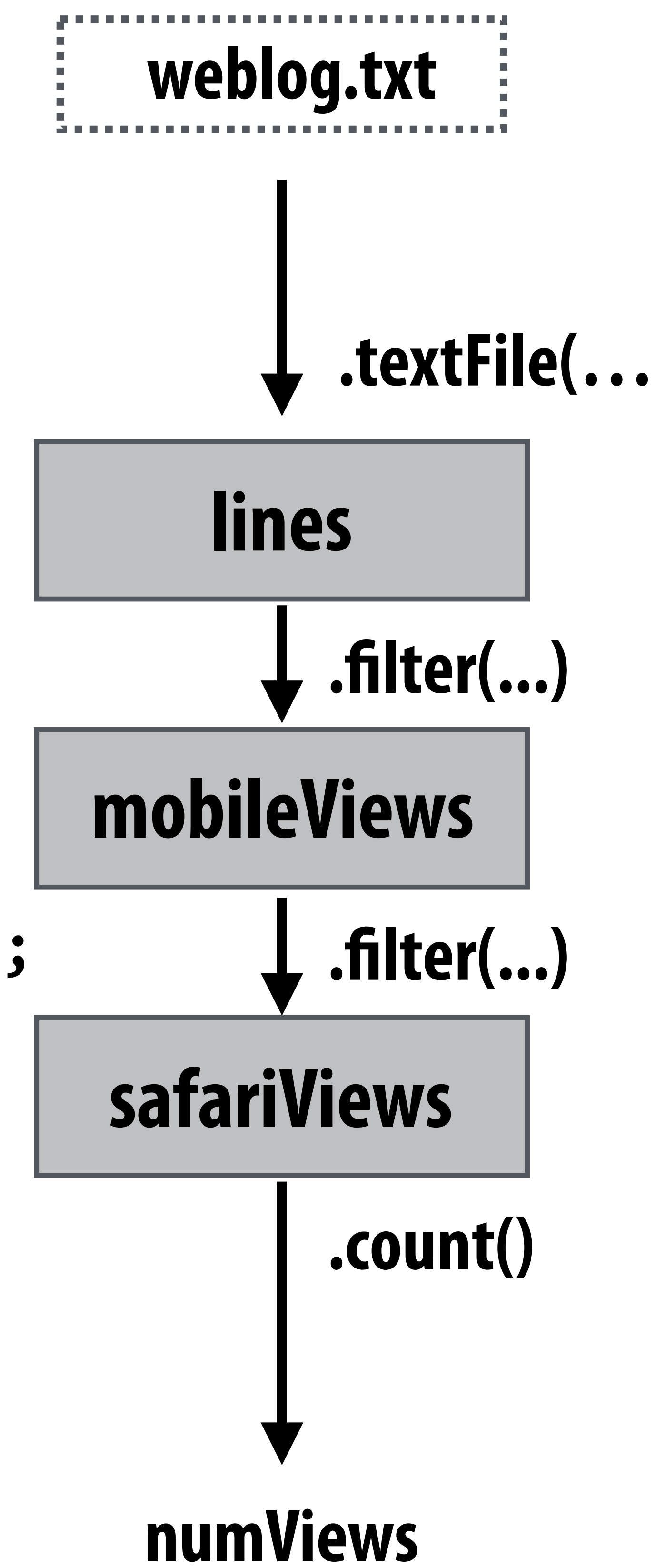
Spark's key programming abstraction:

- Read-only ordered collection of records (immutable)
- RDDs can only be created by deterministic ***transformations*** on data in persistent storage or on existing RDDs
- ***Actions*** on RDDs return data to application

RDDs

```
// create RDD from file system data
var lines = spark.textFile("hdfs://weblog.txt");

// create RDD using filter() transformation on lines
var mobileViews = lines.filter((x: String) => isMobileClient(x));

// another filter() transformation
var safariViews = mobileViews.filter((x: String) => x.contains("Safari"));

// then count number of elements in RDD via count() action
var numViews = safariViews.count();
```

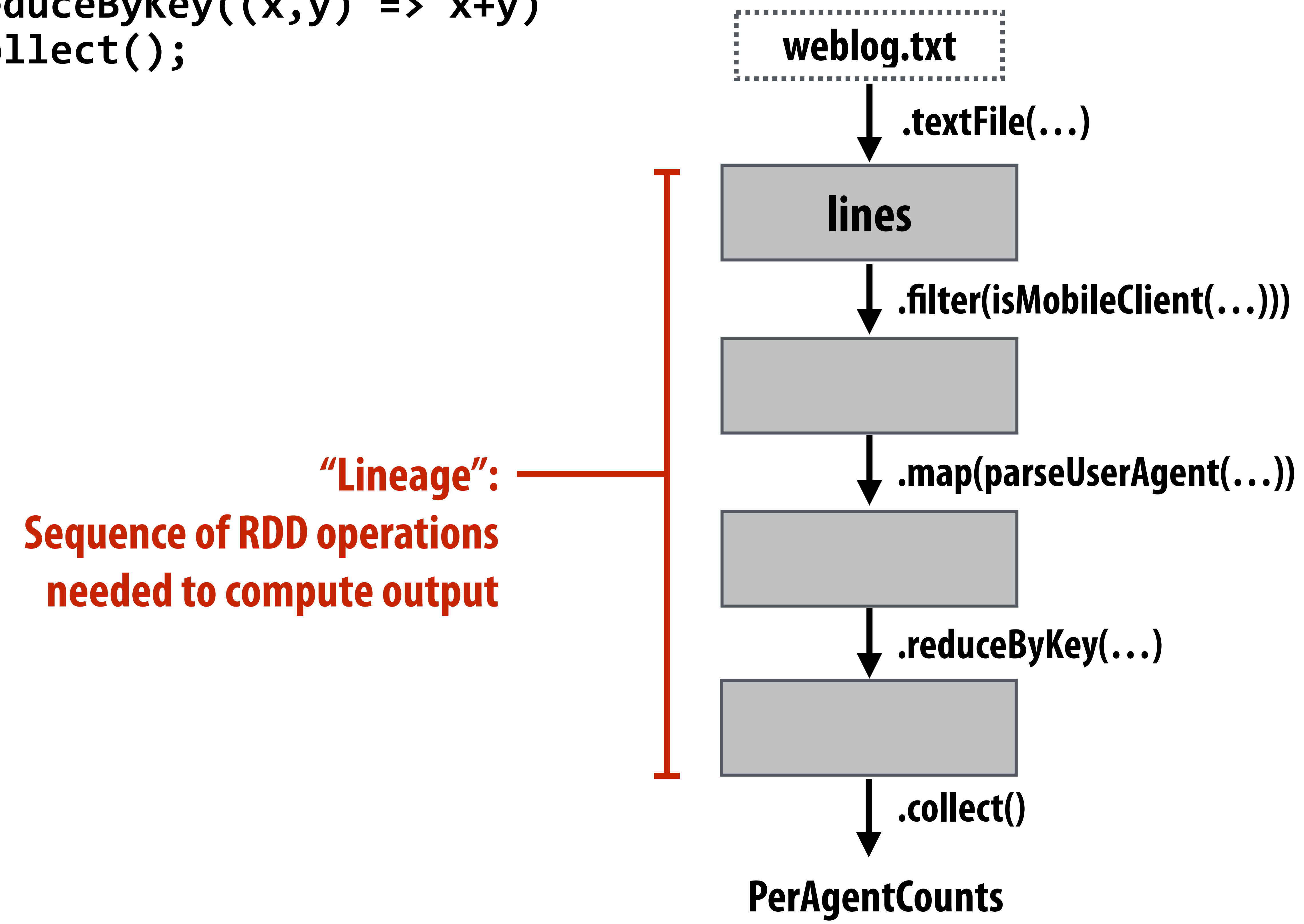
int



Repeating the map-reduce example

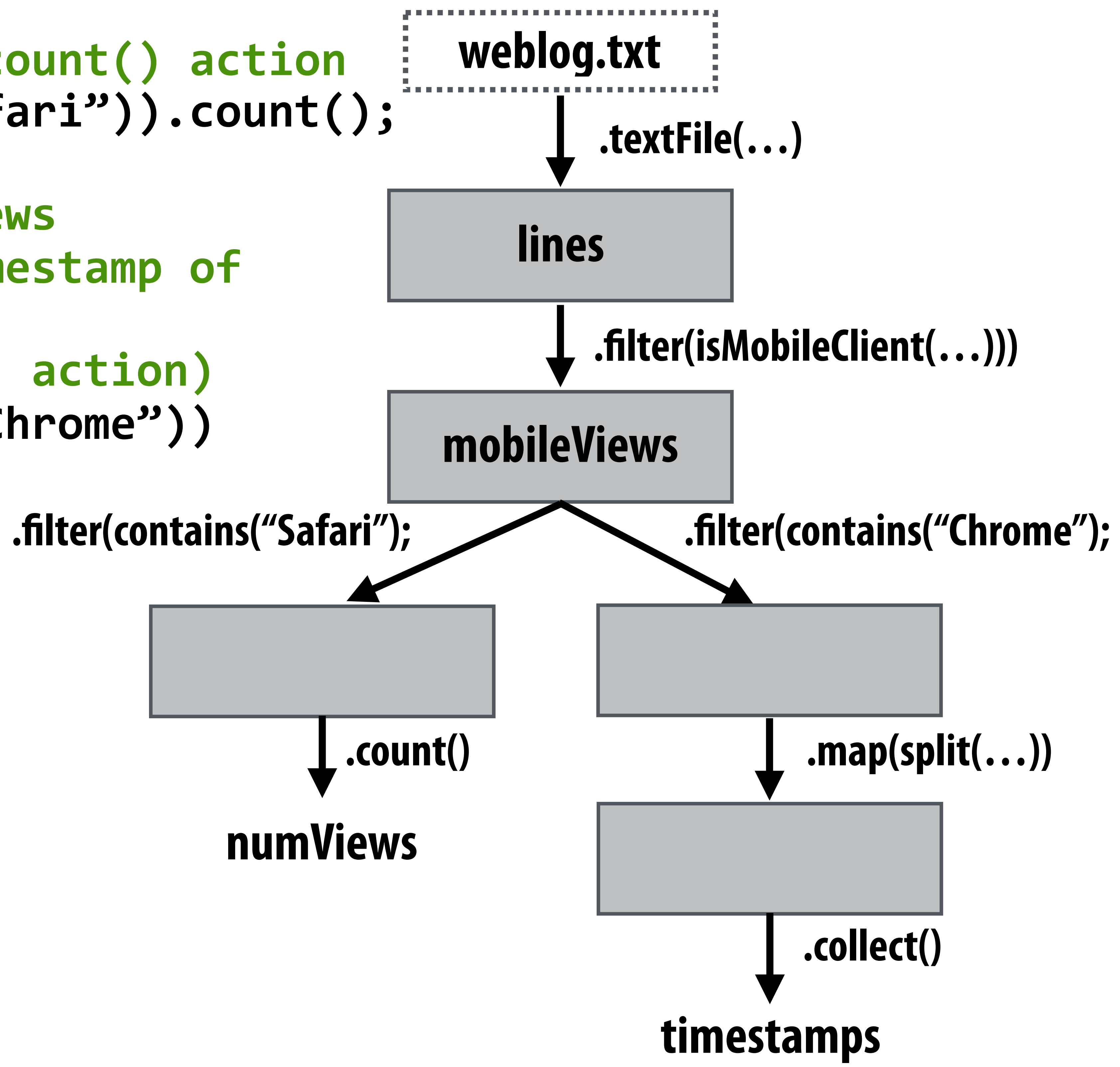
```
// 1. create RDD from file system data
// 2. create RDD with only lines from mobile devices
// 3. create RDD with elements of type (String,Int) from line string
// 4. group elements by key
// 5. call provided reduction function on all keys to count views
var perAgentCounts =  spark.textFile("hdfs://weblog.txt")
                           .filter(x => isMobileDevice(x))
                           .map(x => (parseUserAgent(x),1));
                           .reduceByKey((x,y) => x+y)
                           .collect();
```

Array[String,int]



Another Spark program

```
// create RDD from file system data
var lines = spark.textFile("hdfs://weblog.txt");

// create RDD using filter() transformation on lines
var mobileViews = lines.filter((x: String) => isMobileDevice(x));

// instruct Spark runtime to try to keep mobileViews in memory
mobileViews.persist();

// create a new RDD by filtering mobileViews
// then count number of elements in new RDD via count() action
var numViews = mobileViews.filter(_.contains("Safari")).count();

// 1. create new RDD by filtering only Chrome views
// 2. for each element, split string and take timestamp of
//    page view
// 3. convert RDD to a scalar sequence (collect() action)
var timestamps = mobileViews.filter(_.contains("Chrome"))
                            .map(_.split(" ")(0))
                            .collect();
```



RDD transformations and actions

Transformations: (data parallel operators taking an input RDD to a new RDD)

$$
\begin{array}{rcl}
map(f : \mathrm{T} \Rightarrow \mathrm{U}) & : & \mathrm{RDD[T]} \Rightarrow \mathrm{RDD[U]} \\
filter(f : \mathrm{T} \Rightarrow \mathrm{Bool}) & : & \mathrm{RDD[T]} \Rightarrow \mathrm{RDD[T]} \\
flatMap(f : \mathrm{T} \Rightarrow \mathrm{Seq[U]}) & : & \mathrm{RDD[T]} \Rightarrow \mathrm{RDD[U]} \\
sample(fraction : \mathrm{Float}) & : & \mathrm{RDD[T]} \Rightarrow \mathrm{RDD[T]} \text{ (Deterministic sampling)} \\
groupByKey() & : & \mathrm{RDD[(K, V)]} \Rightarrow \mathrm{RDD[(K, Seq[V])]} \\
reduceByKey(f : (\mathrm{V}, \mathrm{V}) \Rightarrow \mathrm{V}) & : & \mathrm{RDD[(K, V)]} \Rightarrow \mathrm{RDD[(K, V)]} \\
union() & : & (\mathrm{RDD[T]}, \mathrm{RDD[T]}) \Rightarrow \mathrm{RDD[T]} \\
join() & : & (\mathrm{RDD[(K, V)]}, \mathrm{RDD[(K, W)]}) \Rightarrow \mathrm{RDD[(K, (V, W))]} \\
cogroup() & : & (\mathrm{RDD[(K, V)]}, \mathrm{RDD[(K, W)]}) \Rightarrow \mathrm{RDD[(K, (Seq[V], Seq[W]))]} \\
crossProduct() & : & (\mathrm{RDD[T]}, \mathrm{RDD[U]}) \Rightarrow \mathrm{RDD[(T, U)]} \\
mapValues(f : \mathrm{V} \Rightarrow \mathrm{W}) & : & \mathrm{RDD[(K, V)]} \Rightarrow \mathrm{RDD[(K, W)]} \text{ (Preserves partitioning)} \\
sort(c : \mathrm{Comparator[K]}) & : & \mathrm{RDD[(K, V)]} \Rightarrow \mathrm{RDD[(K, V)]} \\
partitionBy(p : \mathrm{Partitioner[K]}) & : & \mathrm{RDD[(K, V)]} \Rightarrow \mathrm{RDD[(K, V)]}
\end{array}
$$

Actions: (provide data back to the "host" application)

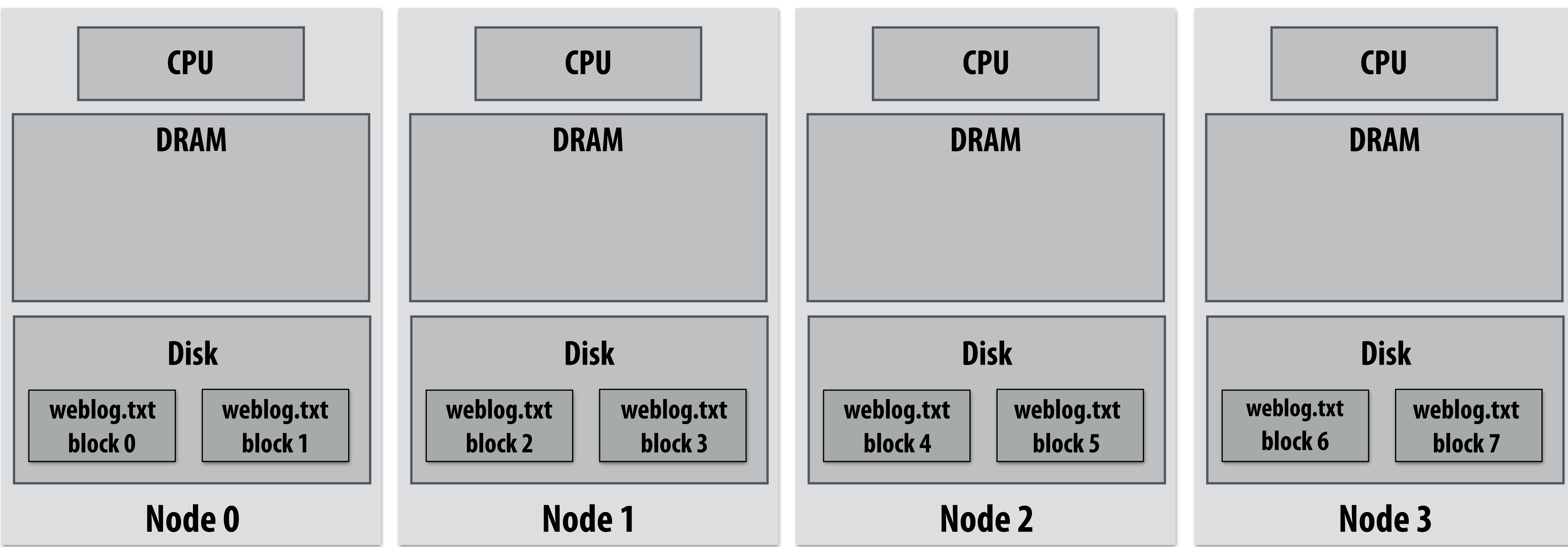
$$
\begin{array}{rcl}
count() & : & \mathrm{RDD[T]} \Rightarrow \mathrm{Long} \\
collect() & : & \mathrm{RDD[T]} \Rightarrow \mathrm{Seq[T]} \\
reduce(f : (\mathrm{T}, \mathrm{T}) \Rightarrow \mathrm{T}) & : & \mathrm{RDD[T]} \Rightarrow \mathrm{T} \\
lookup(k : \mathrm{K}) & : & \mathrm{RDD[(K, V)]} \Rightarrow \mathrm{Seq[V]} \text{ (On hash/range partitioned RDDs)} \\
save(path : \mathrm{String}) & : & \text{Outputs RDD to a storage system, } e.g., \text{ HDFS}
\end{array}
$$

How do we implement RDDs?

In particular, how should they be stored?

```
var lines = spark.textFile("hdfs://weblog.txt");
var lower = lines.map(_.toLower());
var mobileViews = lower.filter(x => isMobileDevice(x));
var howMany = mobileViews.count();
```

Question: should we think of RDD's like arrays?

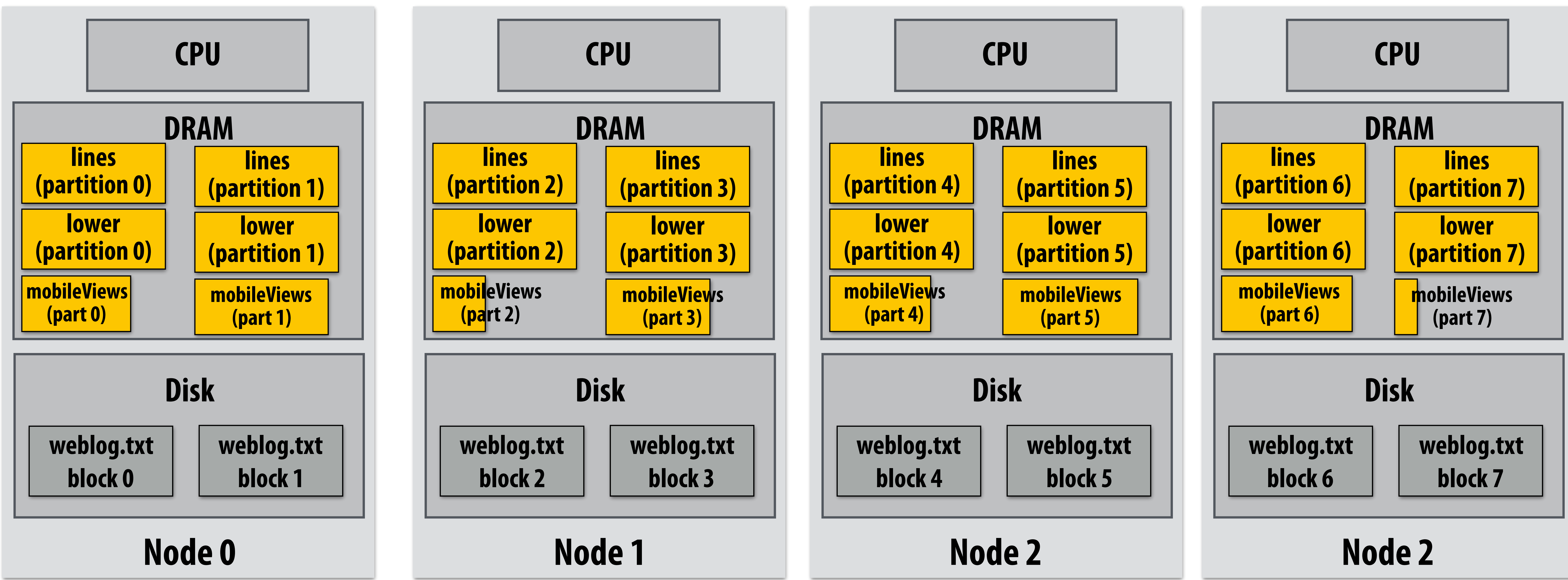


How do we implement RDDs?

In particular, how should they be stored?

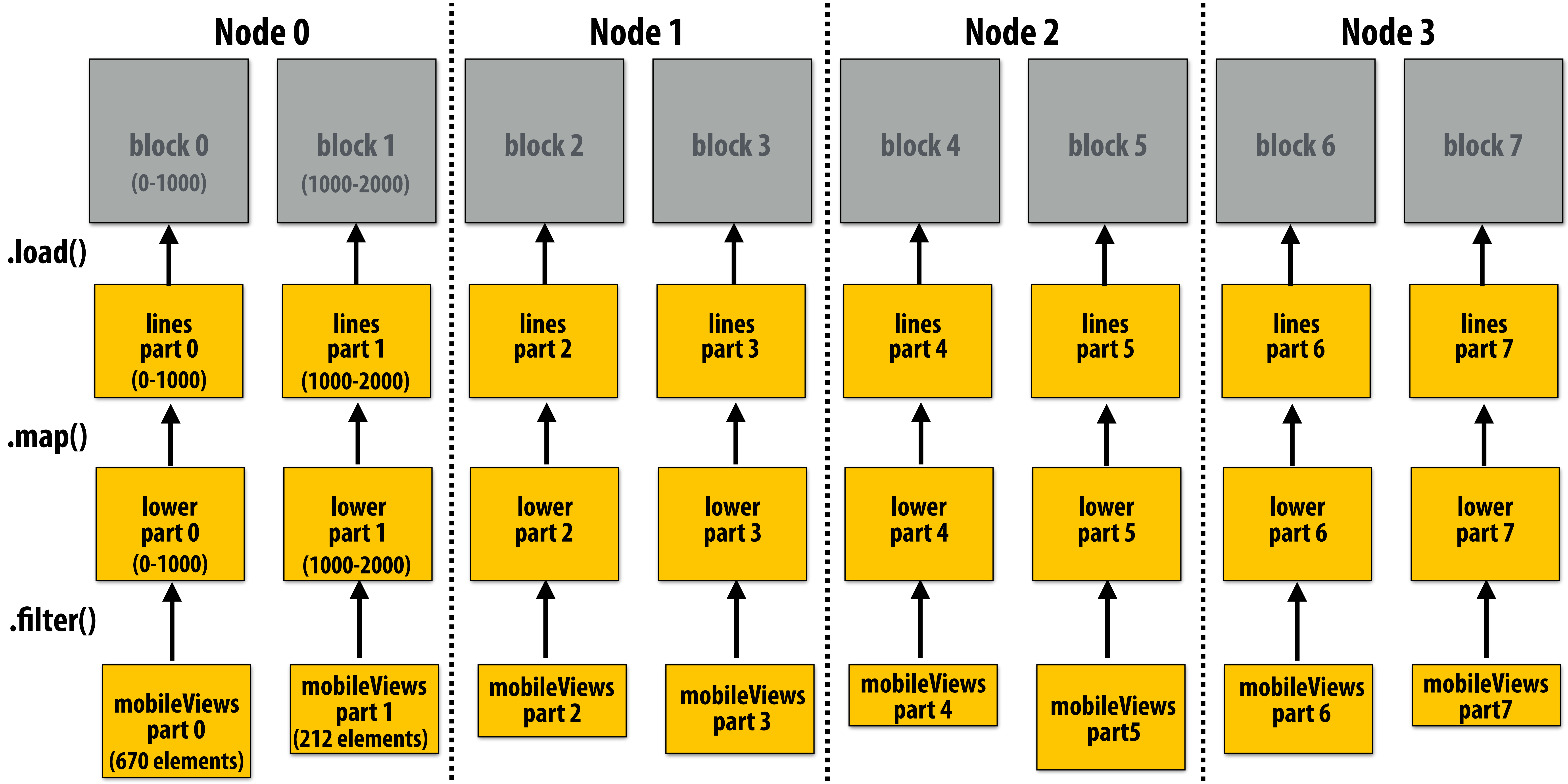
```
var lines = spark.textFile("hdfs://weblog.txt");
var lower = lines.map(_.toLower());
var mobileViews = lower.filter(x => isMobileDevice(x));
var howMany = mobileViews.count();
```

In-memory representation would be huge! (larger than original file on disk)



RDD partitioning and dependencies

```
var lines = spark.textFile("hdfs://weblog.txt");
var lower = lines.map(_.toLower());
var mobileViews = lower.filter(x => isMobileDevice(x));
var howMany = mobileViews.count();
```



Black lines show dependencies between RDD partitions.

Implementing sequence of RDD ops efficiently

```
var lines = spark.textFile("hdfs://weblog.txt");
var lower = lines.map(_.toLower());
var mobileViews = lower.filter(x => isMobileDevice(x));
var howMany = mobileViews.count();
```

Recall "loop fusion" examples from opening slides of lecture

The following code stores only a line of the log file in memory, and only reads input data from disk once ("streaming" solution)

```
int count = 0;
while (inputFile.eof()) {
   string line = inputFile.readLine();
   string lower = line.toLower;
   if (isMobileClient(lower))
     count++;
}
```

A simple interface for RDDs

```
var lines = spark.textFile("hdfs://weblog.txt");
var lower = lines.map(_.toLower());
var mobileViews = lower.filter(x => isMobileDevice(x));
var howMany = mobileViews.count();
```

```
// create RDD by mapping map_func onto
input (parent) RDD
RDD::map(RDD parent, map_func) {
   return new RDDFromMap(parent, map_func);
}

// create RDD by filtering input (parent) RDD
RDD::filter(RDD parent, filter_func) {
   return new RDDFromFilter(parent, filter_func);
}

// create RDD from text file on disk
RDD::textFile(string filename) {
   return new RDDFromTextFile(open(filename));
}

// count action (forces evaluation of RDD)
RDD::count() {
   int count = 0;
   while (hasMoreElements()) {
      var el = next();
      count++;
   }
}
```

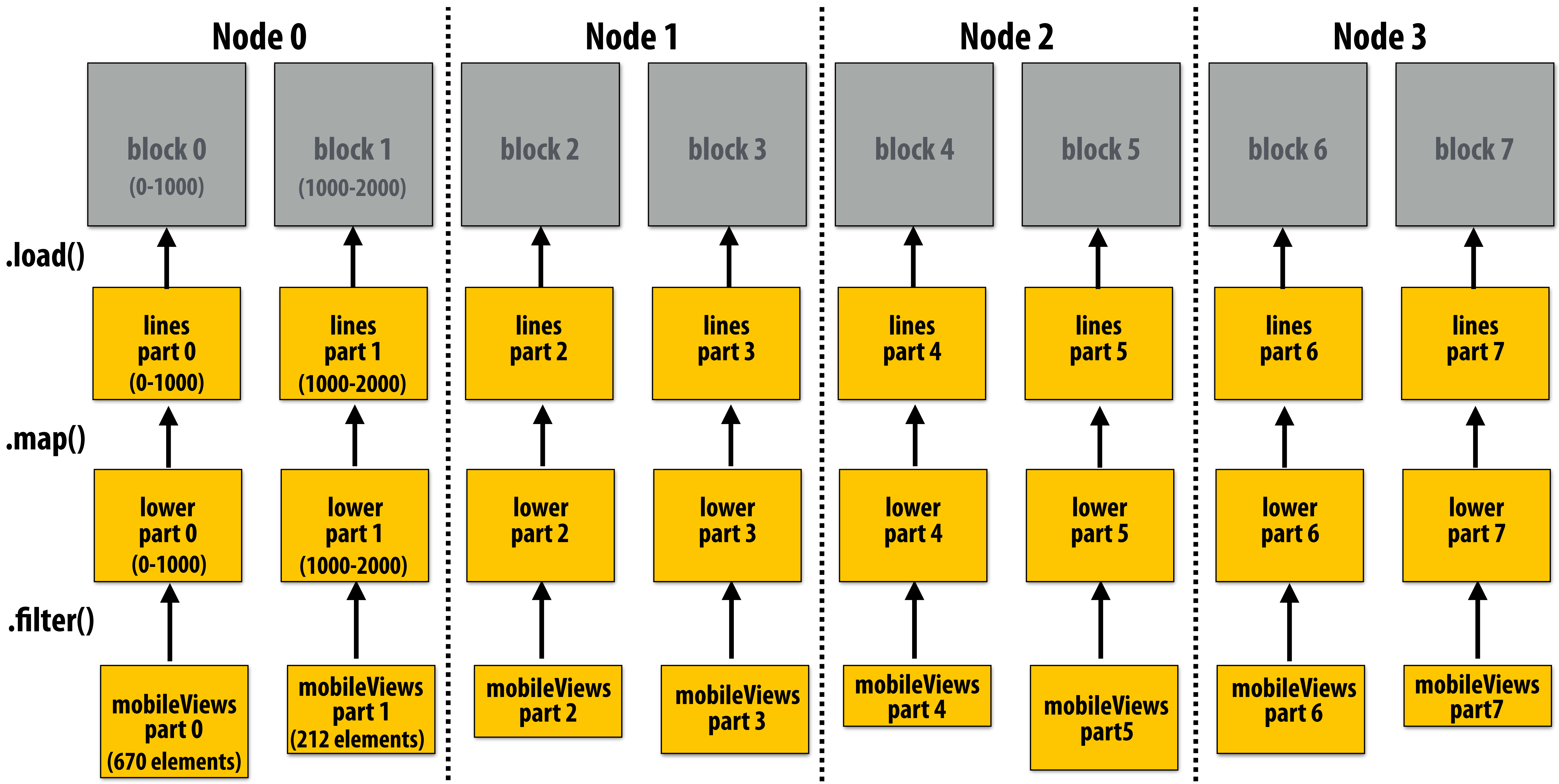
```
RDD::hasMoreElements() {
   parent.hasMoreElements();
}

// overloaded since no parent exists
RDDFromTextFile::hasMoreElements() {
   return !inputFile.eof();
}

RDDFromTextFile::next() {
   return inputFile.readLine();
}

RDDFromMap::next() {
   var el = parent.next();
   return map_func(el);
}

RDDFromFilter::next() {
  while (parent.hasMoreElements()) {
    var el = parent.next();
    if (filter_func(el))
      return el;
}
```

Narrow dependencies

```
var lines = spark.textFile("hdfs://weblog.txt");
var lower = lines.map(_.toLower());
var mobileViews = lower.filter(x => isMobileDevice(x));
var howMany = mobileViews.count();
```

"Narrow dependencies" = each partition of parent RDD referenced by at most one child RDD partition

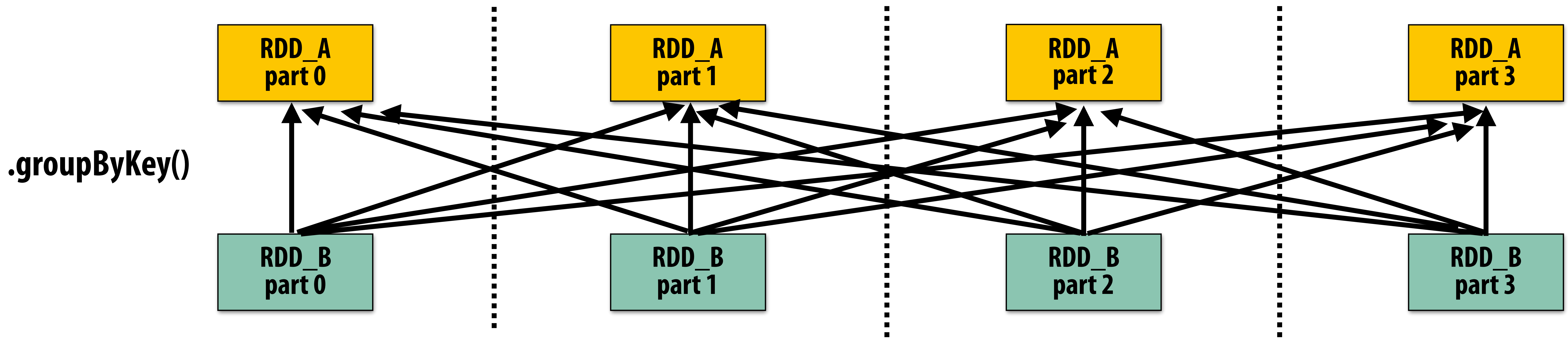
- Allows for fusing of operations (here: can apply map and then filter all at once on input element)
- In this example: no communication between nodes of cluster (communication of one int at end to perform count() reduction)



Wide dependencies

groupByKey: RDD[(K,V)] → RDD[(K,Seq[V])]

"Make a new RDD where each element is a sequence containing all values from the parent RDD with the same key."



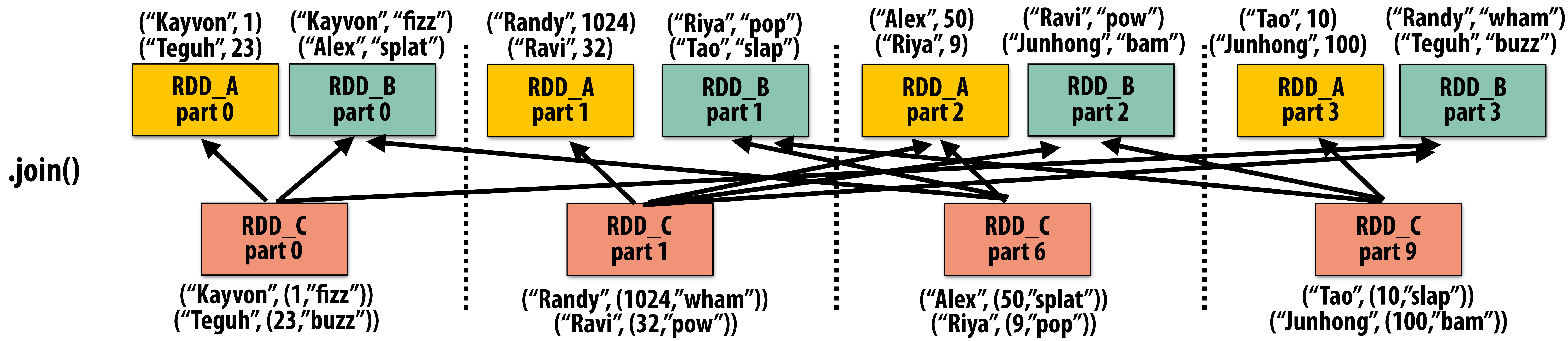
- Wide dependencies = each partition of parent RDD referenced by multiple child RDD partitions
- Challenges:
  - Must compute all of RDD_A before computing RDD_B
    - Example: groupByKey() may induce all-to-all communication as shown above
  - May trigger significant recomputation of ancestor lineage upon node failure
    (I will address resilience in a few slides)

Cost of operations depends on partitioning

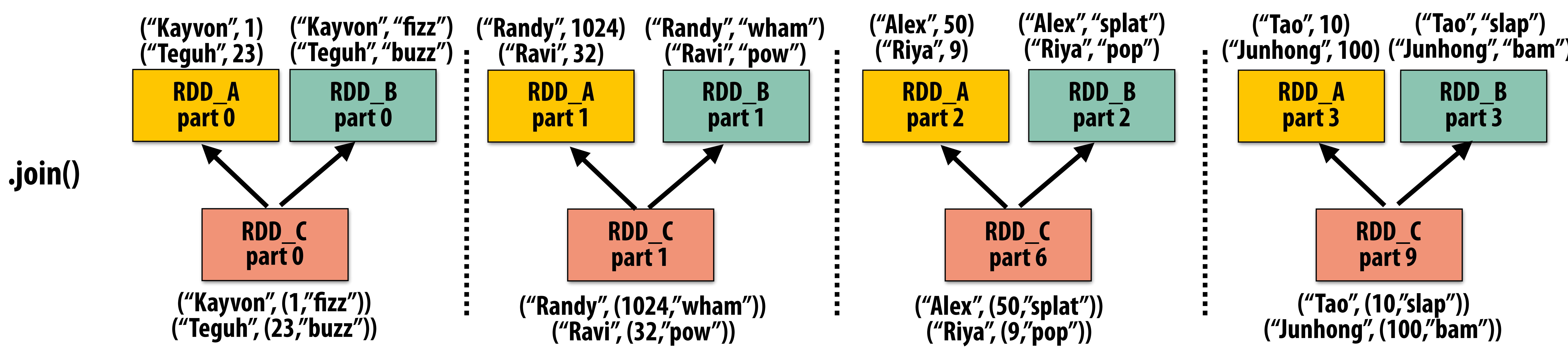
join: RDD[(K,V)], RDD[(K,W)] → RDD[(K,(V,W))]

Assume data in RDD_A and RDD_B are partitioned by key: hash username to partition id

RDD_A and RDD_B have different hash partitions: join creates wide dependencies



RDD_A and RDD_B have same hash partition: join only creates narrow dependencies



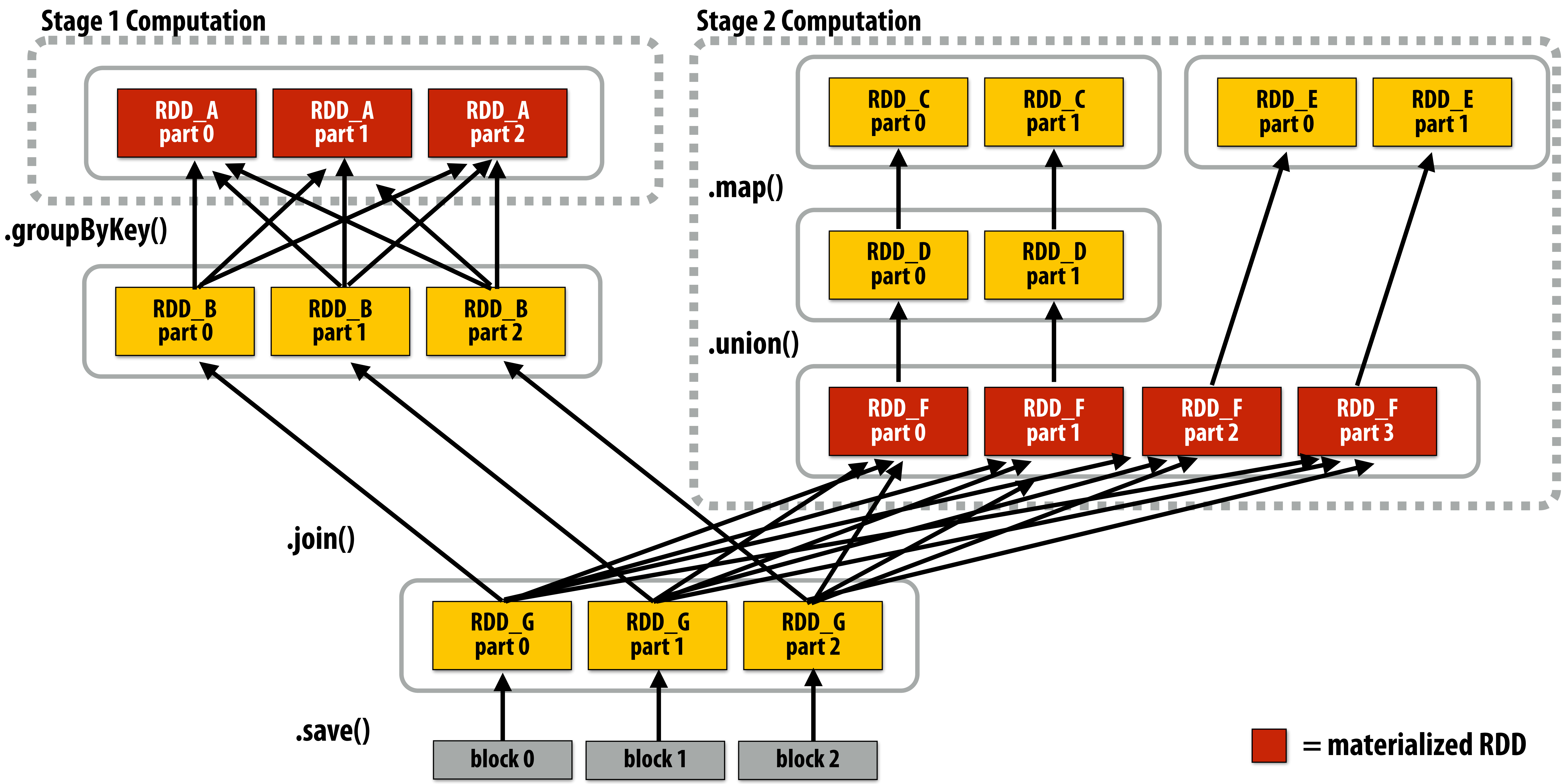
PartitionBy() transformation

- **Inform Spark on how to partition an RDD**
  - **e.g., HashPartitioner, RangePartitioner**

```
// create RDD from file system data
var lines = spark.textFile("hdfs://weblog.txt");
var clientInfo = spark.textFile("hdfs://clientssupported.txt"); // (useragent, "yes"/"no")

// create RDD using filter() transformation on lines
var mobileViews = lines.filter(x => isMobileDevice(x)).map(x => parseUserAgent(x));

// HashPartitioner maps keys to integers
var partitioner = spark.HashPartitioner(100);

// inform Spark of partition
// .persist() also instructs Spark to try to keep dataset in memory
var mobileViewPartitioned = mobileViews.partitionBy(partitioner)
                                        .persist();
var clientInfoPartitioned = clientInfo.partitionBy(partitioner)
                                        .persist();

// join useragents with whether they are supported or not supported
// Note: this join only creates narrow dependencies due to the explicit partitioning above
void joined = mobileViewPartitioned.join(clientInfoPartitioned);
```

- **.persist():**
  - **Inform Spark this RDD's contents should be retained in memory**
  - **.persist(RELIABLE) = store contents in durable storage (like a checkpoint)**

Scheduling Spark computations



Actions (e.g., save()) trigger evaluation of Spark lineage graph.

Stage 1 Computation: do nothing since input already materialized in memory

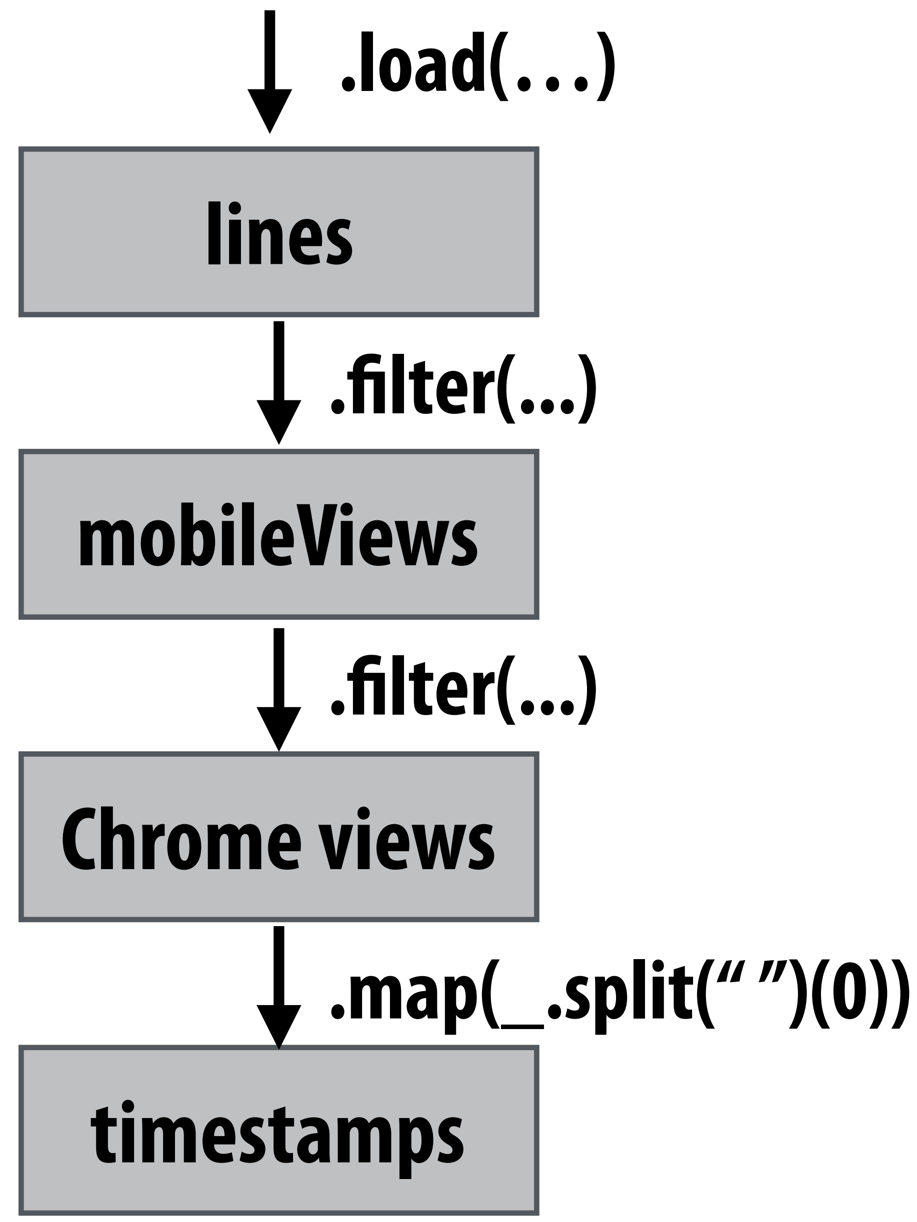
Stage 2 Computation: evaluate map in fused manner, only actually materialize RDD F

Stage 3 Computation: execute join (could stream the operation to disk, do not need to materialize)

Implementing resilience via lineage

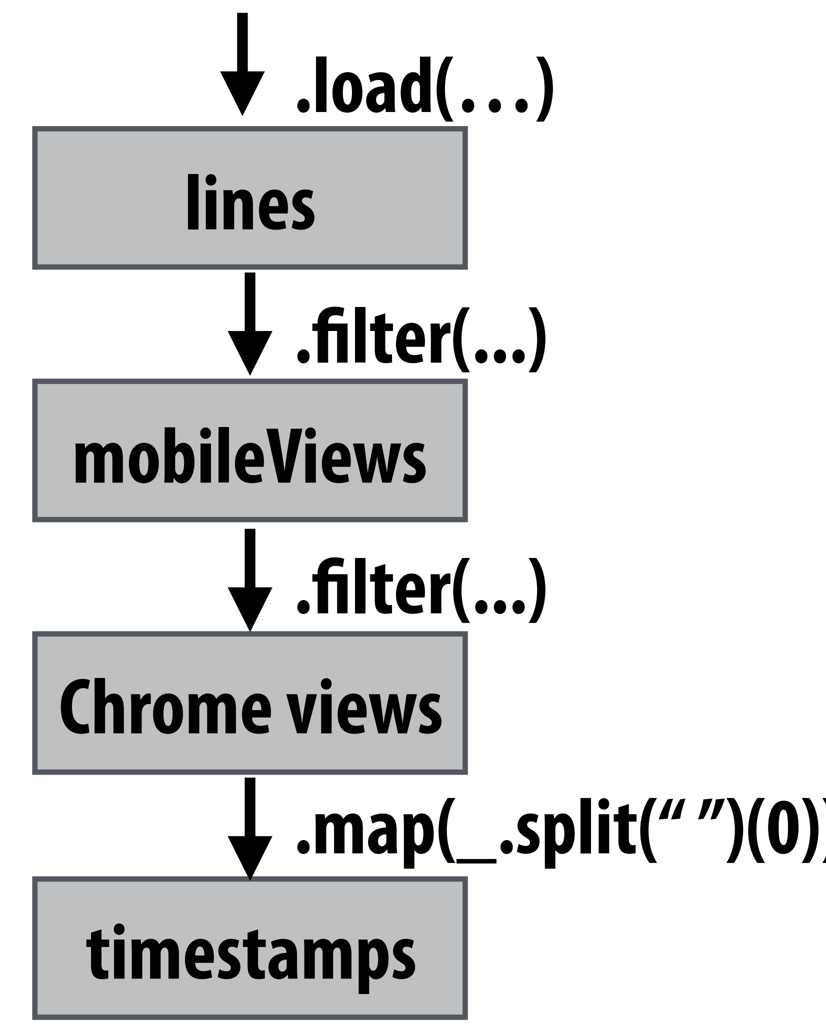
- **RDD transformations are bulk, deterministic, and functional**
  - Implication: runtime can always reconstruct contents of RDD from its lineage (the sequence of transformations used to create it)
  - Lineage is a log of transformations
  - Efficient: since the log records bulk data-parallel operations, overhead of logging is low (compared to logging fine-grained operations, like in a database)

```
// create RDD from file system data
var lines = spark.textFile("hdfs://weblog.txt");

// create RDD using filter() transformation on lines
var mobileViews = lines.filter((x: String) => isMobileDevice(x));

// 1. create new RDD by filtering only Chrome views
// 2. for each element, split string and take timestamp of
//    page view (first element)
// 3. convert RDD To a scalar sequence (collect() action)
var timestamps = mobileView.filter(_.contains("Chrome"))
                           .map(_.split(" ")(0));
```

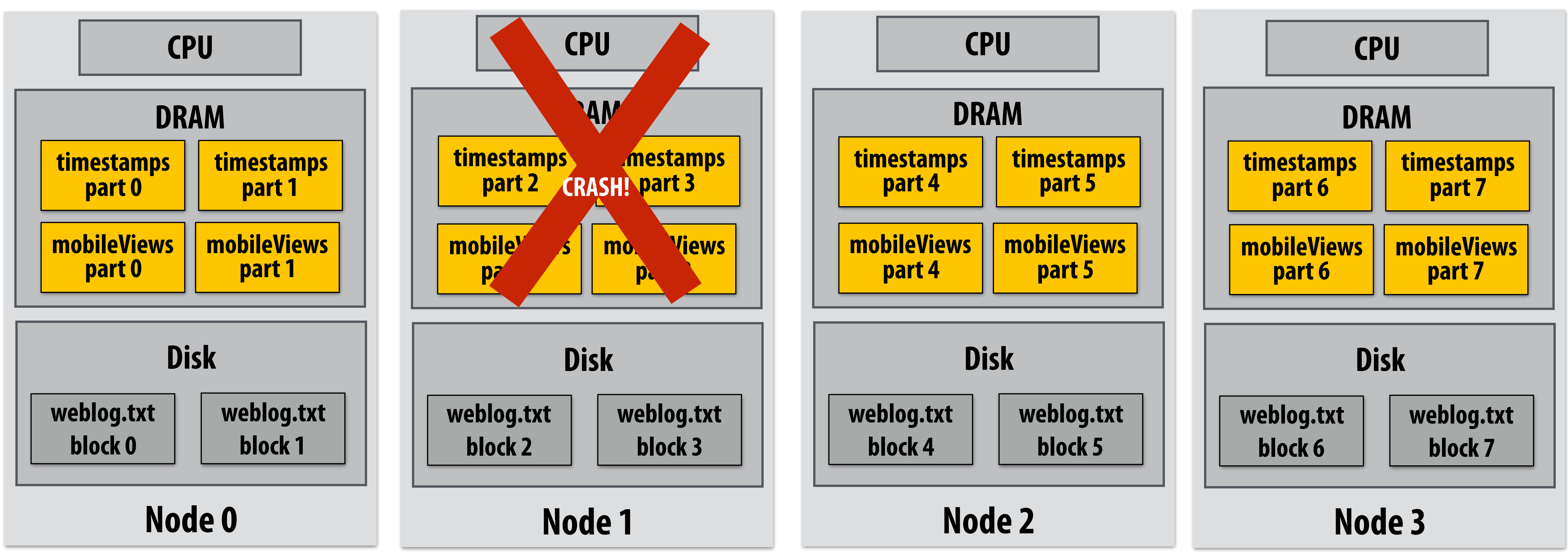


Upon node failure: recompute lost RDD partitions from lineage

```
var lines       = spark.textFile("hdfs://weblog.txt");
var mobileViews = lines.filter((x: String) => isMobileDevice(x));
var timestamps  = mobileView.filter(_.contains("Chrome"))
                            .map(_.split(" ")(0));
```



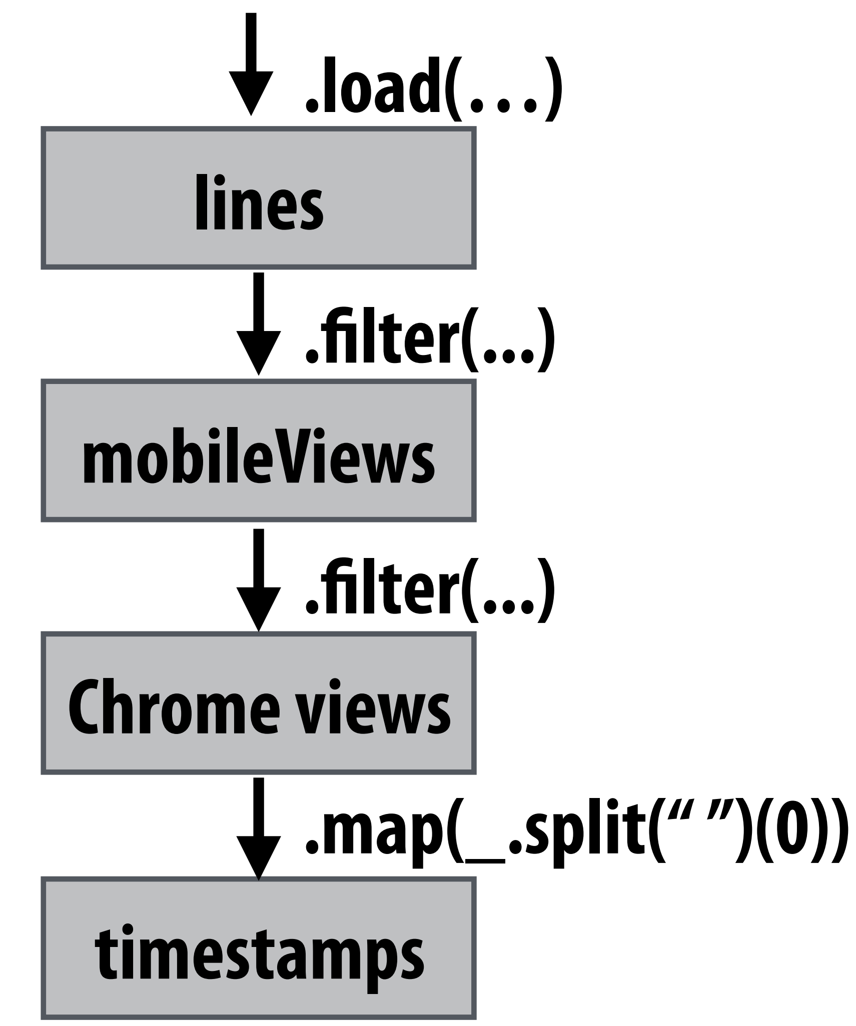
Must reload required subset of data from disk and recompute entire sequence of operations given by lineage to regenerate partitions 2 and 3 of RDD timestamps.



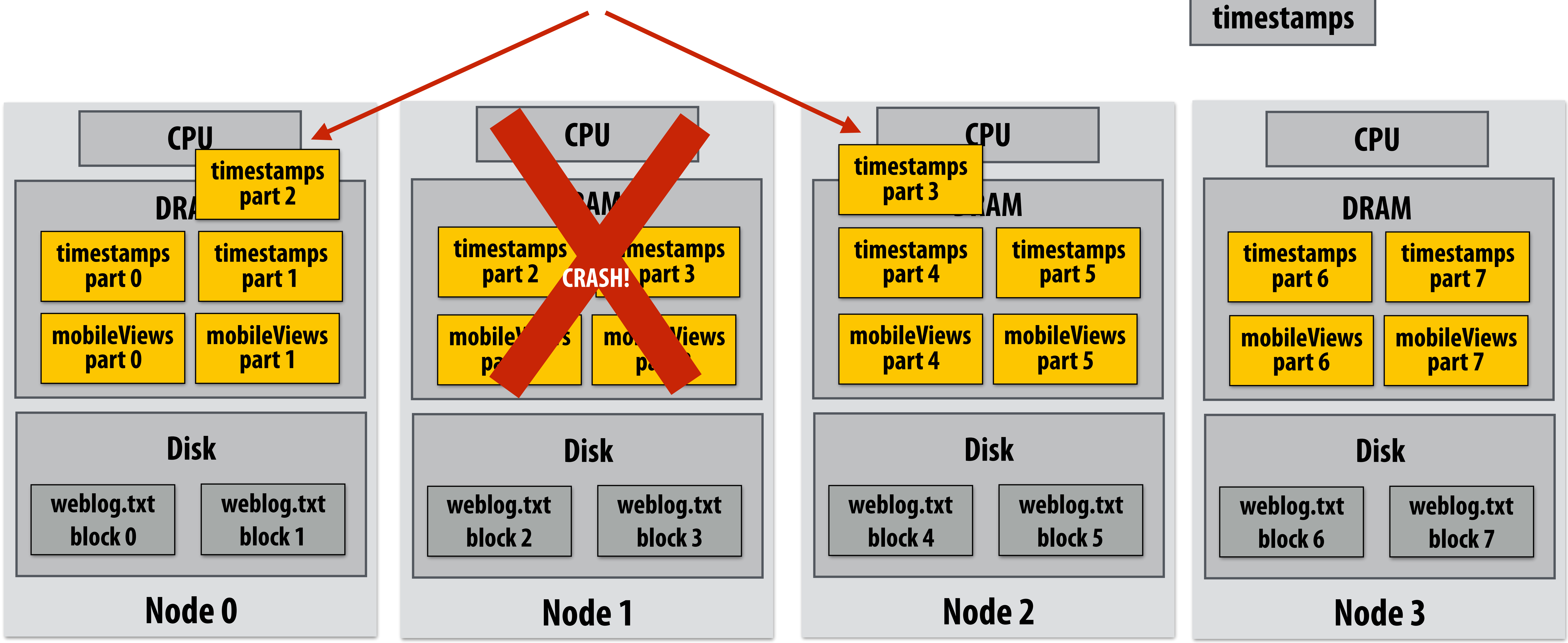
Note: (not shown): file system data is replicated so assume blocks 2 and 3 remain accessible to all nodes

Upon node failure: recompute lost RDD partitions from lineage

```
var lines       = spark.textFile("hdfs://weblog.txt");
var mobileViews = lines.filter((x: String) => isMobileDevice(x));
var timestamps  = mobileView.filter(_.contains("Chrome"))
                            .map(_.split(" ")(0));
```

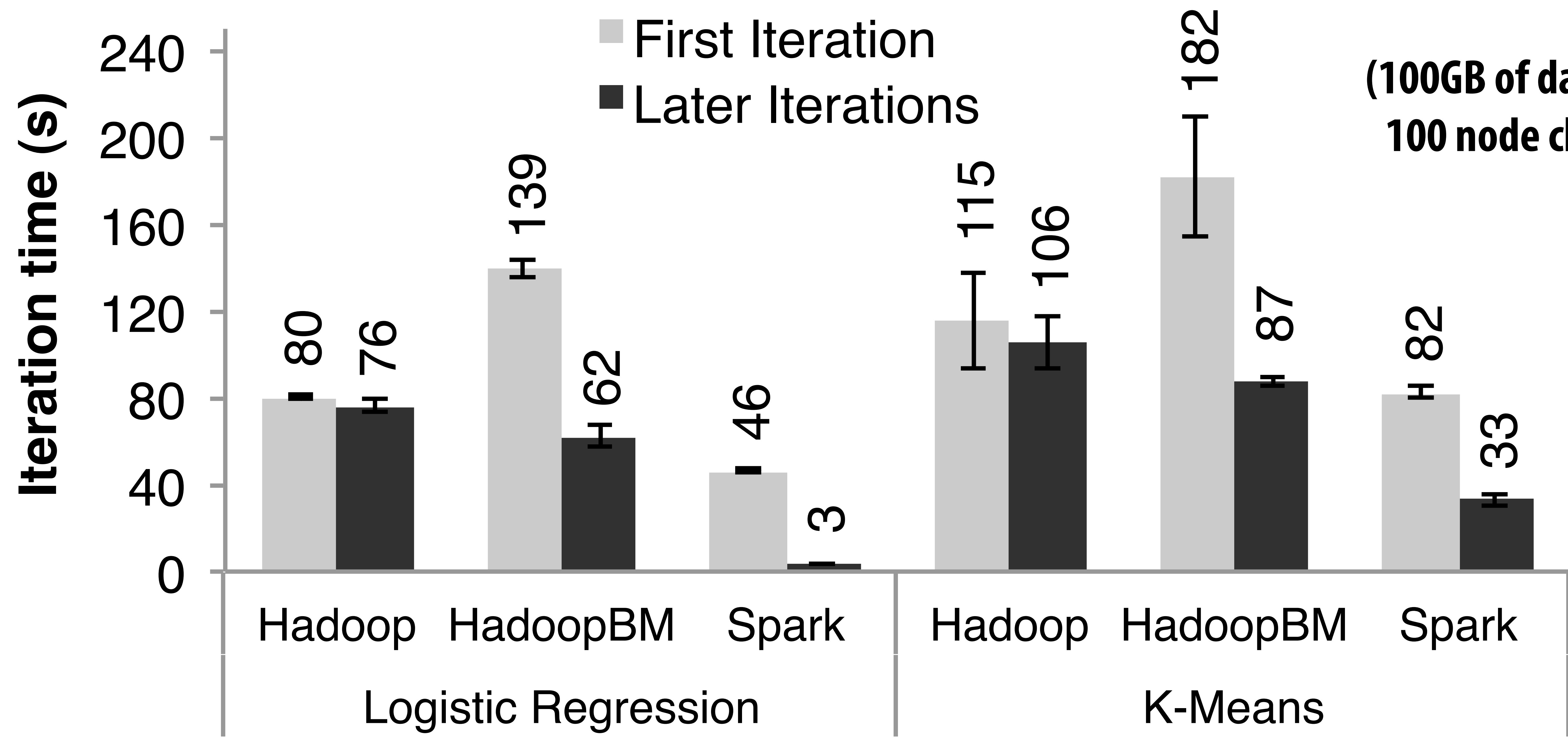


Must reload required subset of data from disk and recompute entire sequence of operations given by lineage to regenerate partitions 2 and 3 of RDD timestamps.



Note: (not shown): file system data is replicated so assume blocks 2 and 3 remain accessible to all nodes

Spark performance



(100GB of data on a 100 node cluster)

HadoopBM = Hadoop Binary In-Memory (convert text input to binary, store in in-memory version of HDFS)

Q. Wait, the baseline parses text input in each iteration of an iterative algorithm? A. Yes.

Anything else puzzling here?

HadoopBM's first iteration is slow because it runs an extra Hadoop job to copy binary form of input data to in memory HDFS

Accessing data from HDFS, even if in memory, has high overhead:

- **Multiple mem copies in file system + a checksum**
- **Conversion from serialized form to Java object**

Caution: "scale out" is not the entire story

- Distributed systems designed for cloud execution address many difficult challenges, and have been instrumental in the explosion of "big-data" computing and large-scale analytics
  - Scale-out parallelism to many machines
  - Resiliency in the face of failures
  - Complexity of managing clusters of machines

- But scale out is not the whole story:

20 Iterations of Page Rank

scalable system	cores	twitter	uk-2007-05
GraphChi [10]	2	3160s	6972s
Stratosphere [6]	16	2250s	-
X-Stream [17]	16	1488s	-
Spark [8]	128	857s	1759s
Giraph [8]	128	596s	1235s
GraphLab [8]	128	249s	833s
GraphX [8]	128	419s	462s
Single thread (SSD)	1	300s	651s
Single thread (RAM)	1	275s	-

name	twitter_rv [11]	uk-2007-05 [4]
nodes	41,652,230	105,896,555
edges	1,468,365,182	3,738,733,648
size	5.76GB	14.72GB

Further optimization of the baseline brought time down to 110s

Vertex order (SSD)	1	300s	651s
Vertex order (RAM)	1	275s	-
Hilbert order (SSD)	1	242s	256s
Hilbert order (RAM)	1	110s	-

["Scalability! At what COST?" McSherry et al. HotOS 2015]

Caution: "scale out" is not the entire story

Label Propagation

[McSherry et al. HotOS 2015]

scalable system	cores	twitter	uk-2007-05
Stratosphere [6]	16	950s	-
X-Stream [17]	16	1159s	-
Spark [8]	128	1784s	$\geq$ 8000s
Giraph [8]	128	200s	$\geq$ 8000s
GraphLab [8]	128	242s	714s
GraphX [8]	128	251s	800s
Single thread (SSD)	1	153s	417s

from McSherry 2015:

"The published work on big data systems has fetishized scalability as the most important feature of a distributed data processing platform. While nearly all such publications detail their system's impressive scalability, few directly evaluate their absolute performance against reasonable benchmarks. To what degree are these systems truly improving performance, as opposed to parallelizing overheads that they themselves introduce?"

COST = "Configuration that Outperforms a Single Thread"

Perhaps surprisingly, many published systems have unbounded COST—i.e., no configuration outperforms the best single-threaded implementation—for all of the problems to which they have been applied.

BID Data Suite (1 GPU accelerated node)

[Canny and Zhao, KDD 13]

Page Rank

System	Graph VxE	Time(s)	Gflops	Procs
Hadoop	?x1.1B	198	0.015	50x8
Spark	40Mx1.5B	97.4	0.03	50x2
Twister	50Mx1.4B	36	0.09	60x4
PowerGraph	40Mx1.4B	3.6	0.8	64x8
BIDMat	60Mx1.4B	6	0.5	1x8
BIDMat+disk	60Mx1.4B	24	0.16	1x8

Latency Dirichlet Allocation (LDA)

System	Docs/hr	Gflops	Procs
Smola[15]	1.6M	0.5	100x8
PowerGraph	1.1M	0.3	64x16
BIDMach	3.6M	30	1x8x1

Performance improvements to Spark

- With increasing DRAM sizes and faster persistent storage (SSD), there is interest in improving the CPU utilization of Spark applications
  - Goal: reduce "COST"

- Efforts looking at adding efficient code generation to Spark ecosystem (e.g., generate SIMD kernels, target accelerators like GPUs, etc.) to close the gap on single node performance
  - RDD storage layouts must change to enable high-performance SIMD processing (e.g., struct of arrays instead of array of structs)
  - See Spark's Project Tungsten, Weld [Palkar Cidr '17], IBM's SparkGPU

- High-performance computing ideas are influencing design of future performance-oriented distributed systems
  - Conversely: the scientific computing community has a lot to learn from the distributed computing community about elasticity and utility computing

Spark summary

- Introduces opaque sequence abstraction (RDD) to encapsulate intermediates of cluster computations (previously… frameworks like Hadoop/MapReduce stored intermediates in the file system)
  - Observation: "files are a poor abstraction for intermediate variables in large-scale data-parallel programs"
  - RDDs are read-only, and created by deterministic data-parallel operators
  - Lineage tracked and used for locality-aware scheduling and fault-tolerance (allows recomputation of partitions of RDD on failure, rather than restore from checkpoint *)
    - Bulk operations allow overhead of lineage tracking (logging) to be low.

- Simple, versatile abstraction upon which many domain-specific distributed computing frameworks are being implemented.
  - See Apache Spark project: spark.apache.org

* Note that .persist(RELIABLE) allows programmer to request checkpointing in long lineage situations.

Modern Spark ecosystem

Compelling feature: enables integration/composition of multiple domain-specific frameworks (since all collections implemented under the hood with RDDs and scheduled using Spark scheduler)

Spark SQL

```
sqlCtx = new HiveContext(sc)
results = sqlCtx.sql(
  "SELECT * FROM people")
names = results.map(lambda p: p.name)
```

Interleave computation and database query
Can apply transformations to RDDs produced by SQL queries

Spark MLlib

```
points = spark.textFile("hdfs://...")
              .map(parsePoint)

model = KMeans.train(points, k=10)
```

Machine learning library build on top of Spark abstractions.

Spark GraphX

```
graph = Graph(vertices, edges)
messages = spark.textFile("hdfs://...")
graph2 = graph.joinVertices(messages) {
  (id, vertex, msg) => ...
}
```

GraphLab-like library built on top of Spark abstractions.