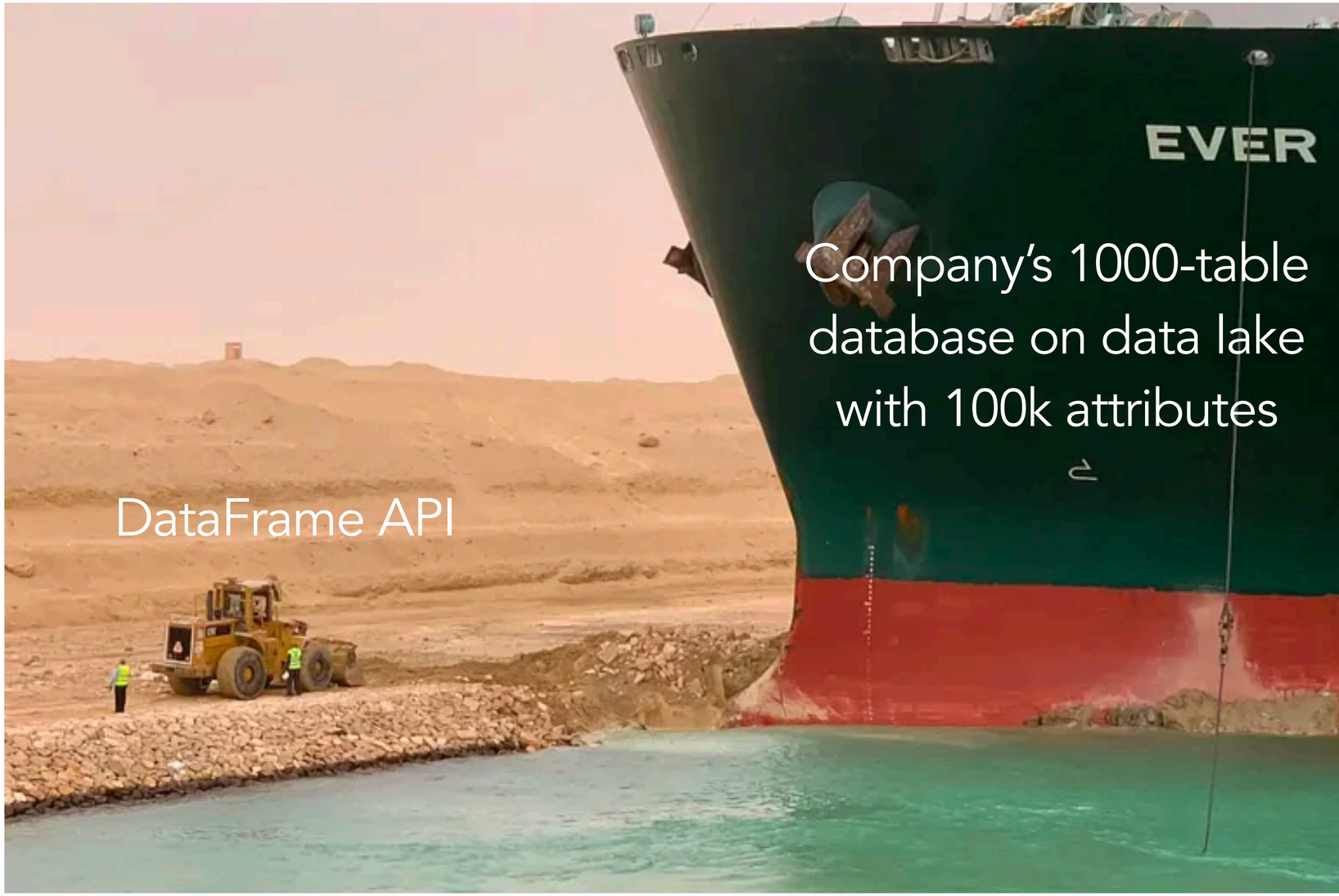
DSC 204a Scalable Data Systems

- Haojian Jin



Meme idea credit: https://datasystemsfun.tumblr.com/

Where are we in the class?

Foundations of Data Systems (2 weeks)

- Digital representation of Data → Computer Organization → Memory hierarchy → Process → Storage

Scaling Distributed Systems (3 weeks)

- Cloud → Network → Distributed storage → Parallelism → Partition and replication

Data Processing and Programming model (5 weeks)

- Data Models evolution → Data encoding evolution → IO & Unix Pipes → Batch processing (MapReduce) → **Stream processing (Spark)**

Today’s topic: Stream Processing

- Overview
-

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

Which program performs better? Program 2

```
float*  A,*B,  *C,  *D,  *E,  *tmp1,   *tmp2;
// assume   arrays are  allocated here
// computeE = D + ((A   + B)*  C)
add(n, A,B, tmp1);
mul(n, tmp1, C,tmp2);
add(n, tmp2, D,    E);



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

Overall arithmetic intensity = 1/3

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

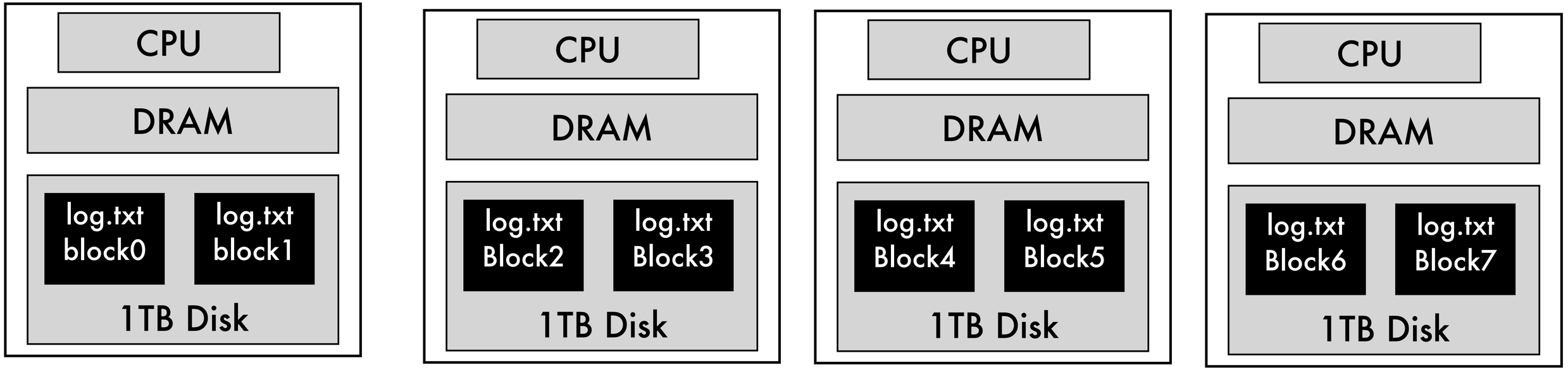
A log of page views on the 418 web site

```
- [05/Apr/2016:22:44:10 -0400] "GET /spring2016content/lectures/16_synchronization/thumbs/slide_012.jpg HTTP/1.1" 200 20186 "http://15418.courses.cs.cmu.edu/spring2016/lecture/synchronization" "Mozilla/5.0 (Macintosh; Intel Ma
- [05/Apr/2016:22:44:10 -0400] "GET /spring2016content/lectures/16_synchronization/thumbs/slide_029.jpg HTTP/1.1" 200 31979 "http://15418.courses.cs.cmu.edu/spring2016/lecture/synchronization" "Mozilla/5.0 (Macintosh; Intel Ma
- [05/Apr/2016:22:44:10 -0400] "GET /spring2016content/lectures/16_synchronization/thumbs/slide_031.jpg HTTP/1.1" 200 8425 "http://15418.courses.cs.cmu.edu/spring2016/lecture/synchronization" "Mozilla/5.0 (Macintosh; Intel Mac
- [05/Apr/2016:22:44:10 -0400] "GET /spring2016content/lectures/16_synchronization/thumbs/slide_035.jpg HTTP/1.1" 200 29266 "http://15418.courses.cs.cmu.edu/spring2016/lecture/synchronization" "Mozilla/5.0 (Macintosh; Intel Ma
- [05/Apr/2016:22:44:10 -0400] "GET /spring2016content/lectures/16_synchronization/thumbs/slide_041.jpg HTTP/1.1" 200 32678 "http://15418.courses.cs.cmu.edu/spring2016/lecture/synchronization" "Mozilla/5.0 (Macintosh; Intel Ma
- [05/Apr/2016:22:44:10 -0400] "GET /spring2016content/lectures/16_synchronization/thumbs/slide_042.jpg HTTP/1.1" 200 32585 "http://15418.courses.cs.cmu.edu/spring2016/lecture/synchronization" "Mozilla/5.0 (Macintosh; Intel Ma
- - [05/Apr/2016:22:44:15 -0400] "GET /spring2016/lecture/snoopimpl/slide_042 HTTP/1.1" 200 3689 "http://15418.courses.cs.cmu.edu/spring2016/lecture/snoopimpl/slide_041" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_3) AppleWeb
- - [05/Apr/2016:22:44:15 -0400] "GET /spring2016content/lectures/12_snoopimpl/images/slide_042.jpg HTTP/1.1" 200 161338 "http://15418.courses.cs.cmu.edu/spring2016/lecture/snoopimpl/slide_042" "Mozilla/5.0 (Macintosh; Intel Ma
- - [05/Apr/2016:22:44:17 -0400] "GET /spring2016/lecture/snoopimpl/slide_041 HTTP/1.1" 200 3093 "http://15418.courses.cs.cmu.edu/spring2016/lecture/snoopimpl/slide_042" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_3) AppleWeb
- [05/Apr/2016:22:44:17 -0400] "GET /spring2016/lecture/synchronization/slide_020 HTTP/1.1" 200 3180 "http://15418.courses.cs.cmu.edu/spring2016/lecture/synchronization" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) AppleWe
- - [05/Apr/2016:22:44:18 -0400] "GET /spring2016/keep_alive HTTP/1.1" 200 957 "http://15418.courses.cs.cmu.edu/spring2016/lecture/basicarch/slide_073" "Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko
- [05/Apr/2016:22:44:18 -0400] "GET /spring2016content/lectures/16_synchronization/images/slide_020.jpg HTTP/1.1" 200 174283 "http://15418.courses.cs.cmu.edu/spring2016/lecture/synchronization/slide_020" "Mozilla/5.0 (Macintos

- [05/Apr/2016:22:44:18 -0400] "GET /spring2016content/profile_pictures/sidwad.jpg HTTP/1.1" 200 34712 "http://15418.courses.cs.cmu.edu/spring2016/lecture/synchronization/slide_020" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_1
- [05/Apr/2016:22:44:18 -0400] "GET /spring2016content/profile_pictures/TomoA.jpg HTTP/1.1" 200 40709 "http://15418.courses.cs.cmu.edu/spring2016/lecture/synchronization/slide_020" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11
- [05/Apr/2016:22:44:18 -0400] "GET /spring2016content/profile_pictures/eknight7.jpg HTTP/1.1" 200 3132 "http://15418.courses.cs.cmu.edu/spring2016/lecture/synchronization/slide_020" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_
- [05/Apr/2016:22:44:18 -0400] "GET /spring2016content/profile_pictures/thomasts.jpg HTTP/1.1" 200 42369 "http://15418.courses.cs.cmu.edu/spring2016/lecture/synchronization/slide_020" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10
- - [05/Apr/2016:22:44:18 -0400] "GET /spring2016/lecture/snoopimpl/slide_040 HTTP/1.1" 200 4985 "http://15418.courses.cs.cmu.edu/spring2016/lecture/snoopimpl/slide_041" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_3) AppleWeb
- - [05/Apr/2016:22:44:19 -0400] "GET /spring2016/lecture/snoopimpl/slide_039 HTTP/1.1" 200 3447 "http://15418.courses.cs.cmu.edu/spring2016/lecture/snoopimpl/slide_040" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_3) AppleWeb
- - [05/Apr/2016:22:44:19 -0400] "GET /spring2016/lecture/snoopimpl/slide_040 HTTP/1.1" 200 4985 "http://15418.courses.cs.cmu.edu/spring2016/lecture/snoopimpl/slide_039" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_3) AppleWeb
- [05/Apr/2016:22:44:21 -0400] "GET /spring2016/users/login HTTP/1.1" 200 2302 "http://15418.courses.cs.cmu.edu/spring2016/lecture/synchronization/slide_020" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) AppleWebKit/601.5.1
- [05/Apr/2016:22:44:26 -0400] "POST /spring2016/users/do_login HTTP/1.1" 302 1061 "http://15418.courses.cs.cmu.edu/spring2016/users/login" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) AppleWebKit/601.5.17 (KHTML, like Ge
- [05/Apr/2016:22:44:26 -0400] "GET /spring2016/ HTTP/1.1" 200 4767 "http://15418.courses.cs.cmu.edu/spring2016/users/login" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) AppleWebKit/601.5.17 (KHTML, like Gecko) Version/9.1
- [05/Apr/2016:22:44:26 -0400] "GET /spring2016content/profile_pictures/cmusam.jpg HTTP/1.1" 200 42983 "http://15418.courses.cs.cmu.edu/spring2016/" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) AppleWebKit/601.5.17 (KHTML,
- [05/Apr/2016:22:44:30 -0400] "GET /spring2016/lectures HTTP/1.1" 200 6322 "http://15418.courses.cs.cmu.edu/spring2016/" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) AppleWebKit/601.5.17 (KHTML, like Gecko) Version/9.1 Sa
- [05/Apr/2016:22:44:33 -0400] "GET /spring2016/lecture/synchronization HTTP/1.1" 200 2871 "http://15418.courses.cs.cmu.edu/spring2016/lectures" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) AppleWebKit/601.5.17 (KHTML, lik
- [05/Apr/2016:22:44:35 -0400] "GET /spring2013content/lectures/03_progmodels/images/slide_032.png HTTP/1.1" 304 189 "-" "Mozilla/5.0 (compatible; YandexImages/3.0; +http://yandex.com/bots)"
- [05/Apr/2016:22:44:38 -0400] "GET /spring2016/lecture/synchronization/slide_020 HTTP/1.1" 200 3852 "http://15418.courses.cs.cmu.edu/spring2016/lecture/synchronization" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) AppleWe
- [05/Apr/2016:22:45:00 -0400] "GET /spring2013/article/26 HTTP/1.1" 200 5900 "http://www.google.co.in/search?ie=UTF-8&q=split+transaction+bus&revid=112973050&sa=X&ved=0ahUKEwio0PfG_vjLAhVinIMKHQO5AdYQ1QIIBQ" "UCWEB/2.0 (Java;
- [05/Apr/2016:22:45:01 -0400] "GET /spring2013/assets/js/15418_common.js HTTP/1.1" 200 425 "http://15418.courses.cs.cmu.edu/spring2013/article/26" "UCWEB/2.0 (Java; U; MIDP-2.0; Nokia203/20.37) U2/1.0.0 UCBrowser/8.7.0.218 U2/
- [05/Apr/2016:22:45:01 -0400] "GET /spring2013/assets/third_party/jquery/timeago/jquery.timeago.js HTTP/1.1" 200 2026 "http://15418.courses.cs.cmu.edu/spring2013/article/26" "UCWEB/2.0 (Java; U; MIDP-2.0; Nokia203/20.37) U2/1.
- [05/Apr/2016:22:45:01 -0400] "GET /spring2013/assets/third_party/jquery/cookie/jquery.cookie.js HTTP/1.1" 200 1189 "http://15418.courses.cs.cmu.edu/spring2013/article/26" "UCWEB/2.0 (Java; U; MIDP-2.0; Nokia203/20.37) U2/1.0.
- [05/Apr/2016:22:45:01 -0400] "GET /spring2013/assets/third_party/date/date.js HTTP/1.1" 200 7628 "http://15418.courses.cs.cmu.edu/spring2013/article/26" "UCWEB/2.0 (Java; U; MIDP-2.0; Nokia203/20.37) U2/1.0.0 UCBrowser/8.7.0.
- [05/Apr/2016:22:45:01 -0400] "GET /spring2013/assets/third_party/codemirror-3.0/mode/markdown/markdown.js HTTP/1.1" 200 4018 "http://15418.courses.cs.cmu.edu/spring2013/article/26" "UCWEB/2.0 (Java; U; MIDP-2.0; Nokia203/20.3
- [05/Apr/2016:22:45:01 -0400] "GET /spring2013/assets/third_party/google-code-prettify/prettify.js HTTP/1.1" 200 6379 "http://15418.courses.cs.cmu.edu/spring2013/article/26" "UCWEB/2.0 (Java; U; MIDP-2.0; Nokia203/20.37) U2/1.
- [05/Apr/2016:22:45:01 -0400] "GET /spring2013/assets/third_party/jquery/tmpl/jquery.tmpl.min.js HTTP/1.1" 200 3155 "http://15418.courses.cs.cmu.edu/spring2013/article/26" "UCWEB/2.0 (Java; U; MIDP-2.0; Nokia203/20.37) U2/1.0.
- [05/Apr/2016:22:45:01 -0400] "GET /spring2013/assets/css/main.css HTTP/1.1" 200 3368 "http://15418.courses.cs.cmu.edu/spring2013/article/26" "UCWEB/2.0 (Java; U; MIDP-2.0; Nokia203/20.37) U2/1.0.0 UCBrowser/8.7.0.218 U2/1.0.
- [05/Apr/2016:22:45:01 -0400] "GET /spring2013/assets/third_party/jquery/1.8.3/jquery.min.js HTTP/1.1" 200 33789 "http://15418.courses.cs.cmu.edu/spring2013/article/26" "UCWEB/2.0 (Java; U; MIDP-2.0; Nokia203/20.37) U2/1.0.0 U
- [05/Apr/2016:22:45:01 -0400] "GET /spring2013/assets/third_party/codemirror-3.0/lib/codemirror.css HTTP/1.1" 200 2319 "http://15418.courses.cs.cmu.edu/spring2013/article/26" "UCWEB/2.0 (Java; U; MIDP-2.0; Nokia203/20.37) U2/1
- [05/Apr/2016:22:45:01 -0400] "GET /spring2013/assets/third_party/google-code-prettify/prettify.css HTTP/1.1" 200 660 "http://15418.courses.cs.cmu.edu/spring2013/article/26" "UCWEB/2.0 (Java; U; MIDP-2.0; Nokia203/20.37) U2/1.
- [05/Apr/2016:22:45:01 -0400] "GET /spring2013/assets/js/main.js HTTP/1.1" 200 1512 "http://15418.courses.cs.cmu.edu/spring2013/article/26" "UCWEB/2.0 (Java; U; MIDP-2.0; Nokia203/20.37) U2/1.0.0 UCBrowser/8.7.0.218 U2/1.0.0 M
- [05/Apr/2016:22:45:01 -0400] "GET /spring2013/assets/third_party/codemirror-3.0/lib/codemirror.js HTTP/1.1" 200 47855 "http://15418.courses.cs.cmu.edu/spring2013/article/26" "UCWEB/2.0 (Java; U; MIDP-2.0; Nokia203/20.37) U2/1
- [05/Apr/2016:22:45:01 -0400] "GET /spring2013/assets/js/comments.js HTTP/1.1" 200 2413 "http://15418.courses.cs.cmu.edu/spring2013/article/26" "UCWEB/2.0 (Java; U; MIDP-2.0; Nokia203/20.37) U2/1.0.0 UCBrowser/8.7.0.218 U2/1.0
- [05/Apr/2016:22:45:01 -0400] "GET /spring2013//assets/images/favicon/dragon.png HTTP/1.1" 200 3145 "-" "UCWEB/2.0 (Java; U; MIDP-2.0; Nokia203/20.37) U2/1.0.0 UCBrowser/8.7.0.218 U2/1.0.0 Mobile"
- [05/Apr/2016:22:45:01 -0400] "GET /spring2013content/article_images/26_3.jpg HTTP/1.1" 200 28441 "http://15418.courses.cs.cmu.edu/spring2013/article/26" "UCWEB/2.0 (Java; U; MIDP-2.0; Nokia203/20.37) U2/1.0.0 UCBrowser/8.7.0.
- [05/Apr/2016:22:45:01 -0400] "GET /spring2013content/article_images/26_2.jpg HTTP/1.1" 200 25683 "http://15418.courses.cs.cmu.edu/spring2013/article/26" "UCWEB/2.0 (Java; U; MIDP-2.0; Nokia203/20.37) U2/1.0.0 UCBrowser/8.7.0.
- [05/Apr/2016:22:45:01 -0400] "GET /spring2013content/article_images/26_4.jpg HTTP/1.1" 200 38414 "http://15418.courses.cs.cmu.edu/spring2013/article/26" "UCWEB/2.0 (Java; U; MIDP-2.0; Nokia203/20.37) U2/1.0.0 UCBrowser/8.7.0.
- [05/Apr/2016:22:45:01 -0400] "GET /spring2013content/profile_pictures/lazyplus.jpg HTTP/1.1" 200 40708 "http://15418.courses.cs.cmu.edu/spring2013/article/26" "UCWEB/2.0 (Java; U; MIDP-2.0; Nokia203/20.37) U2/1.0.0 UCBrowser/
- - [05/Apr/2016:22:45:10 -0400] "GET /spring2016/keep_alive HTTP/1.1" 200 957 "http://15418.courses.cs.cmu.edu/spring2016/article/9" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/49.0.2623
- - [05/Apr/2016:22:46:31 -0400] "GET / HTTP/1.1" 302 564 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_3) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/49.0.2623.110 Safari/537.36"
- - [05/Apr/2016:22:46:31 -0400] "GET /spring2016 HTTP/1.1" 301 584 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_3) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/49.0.2623.110 Safari/537.36"
- - [05/Apr/2016:22:46:31 -0400] "GET /spring2016/ HTTP/1.1" 200 5254 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_3) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/49.0.2623.110 Safari/537.36"
- [05/Apr/2016:22:46:40 -0400] "GET /spring2016/lecture/lockfree/slide_028 HTTP/1.1" 200 4008 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
- - [05/Apr/2016:22:46:41 -0400] "GET /spring2016/keep_alive HTTP/1.1" 200 958 "http://15418.courses.cs.cmu.edu/spring2016/article/14" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_3) AppleWebKit/537.36 (KHTML, like Gecko) Chr
- [05/Apr/2016:22:46:51 -0400] "GET /spring2016/keep_alive HTTP/1.1" 200 999 "http://15418.courses.cs.cmu.edu/spring2016/article/14" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/49.0.2623.
- [05/Apr/2016:22:47:00 -0400] "GET /spring2014content/lectures/24_memory/images/slide_014.png HTTP/1.1" 304 190 "-" "Mozilla/5.0 (compatible; YandexImages/3.0; +http://yandex.com/bots)"
- [05/Apr/2016:22:47:25 -0400] "GET /spring2015/article/11 HTTP/1.1" 200 6180 "https://www.google.com/" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/49.0.2623.110 Safari/537.36"
- [05/Apr/2016:22:47:26 -0400] "GET /spring2015/assets/third_party/jquery/timeago/jquery.timeago.js HTTP/1.1" 200 2026 "http://15418.courses.cs.cmu.edu/spring2015/article/11" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) App
- [05/Apr/2016:22:47:26 -0400] "GET /spring2015/assets/third_party/jquery/1.8.3/jquery.min.js HTTP/1.1" 200 33788 "http://15418.courses.cs.cmu.edu/spring2015/article/11" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) AppleWeb
- [05/Apr/2016:22:47:26 -0400] "GET /spring2015/assets/third_party/codemirror-3.0/lib/codemirror.css HTTP/1.1" 200 2319 "http://15418.courses.cs.cmu.edu/spring2015/article/11" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) Ap
- [05/Apr/2016:22:47:26 -0400] "GET /spring2015/assets/third_party/google-code-prettify/prettify.css HTTP/1.1" 200 660 "http://15418.courses.cs.cmu.edu/spring2015/article/11" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) App
- [05/Apr/2016:22:47:26 -0400] "GET /spring2015/assets/css/main.css HTTP/1.1" 200 4011 "http://15418.courses.cs.cmu.edu/spring2015/article/11" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) AppleWebKit/537.36 (KHTML, like Ge
- [05/Apr/2016:22:47:26 -0400] "GET /spring2015/assets/third_party/jquery/cookie/jquery.cookie.js HTTP/1.1" 200 1189 "http://15418.courses.cs.cmu.edu/spring2015/article/11" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) Apple
- [05/Apr/2016:22:47:26 -0400] "GET /spring2015/assets/third_party/codemirror-3.0/lib/codemirror.js HTTP/1.1" 200 47854 "http://15418.courses.cs.cmu.edu/spring2015/article/11" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) Ap
- [05/Apr/2016:22:47:26 -0400] "GET /spring2015/assets/third_party/codemirror-3.0/mode/markdown/markdown.js HTTP/1.1" 200 4017 "http://15418.courses.cs.cmu.edu/spring2015/article/11" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_1
- [05/Apr/2016:22:47:26 -0400] "GET /spring2015/assets/third_party/google-code-prettify/prettify.js HTTP/1.1" 200 6378 "http://15418.courses.cs.cmu.edu/spring2015/article/11" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) App
- [05/Apr/2016:22:47:26 -0400] "GET /spring2015/assets/js/main.js HTTP/1.1" 200 806 "http://15418.courses.cs.cmu.edu/spring2015/article/11" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) AppleWebKit/537.36 (KHTML, like Gecko)
- [05/Apr/2016:22:47:26 -0400] "GET /spring2015/assets/js/comments.js HTTP/1.1" 200 3779 "http://15418.courses.cs.cmu.edu/spring2015/article/11" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) AppleWebKit/537.36 (KHTML, like
- [05/Apr/2016:22:47:29 -0400] "GET /spring2015/assets/images/favicon/dragon.png HTTP/1.1" 200 3145 "http://15418.courses.cs.cmu.edu/spring2015/article/11" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_4) AppleWebKit/537.36 (KH
- - [05/Apr/2016:22:47:41 -0400] "GET /spring2015content/lectures/21_heterogeneity/thumbs/slide_037.png HTTP/1.1" 304 133 "-" "Mozilla/4.0 (compatible;)"
- [05/Apr/2016:22:48:50 -0400] "GET /spring2016content/article_images/3_.jpg HTTP/1.1" 200 65173 "https://www.google.com/" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_0) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/49.0.262
- [05/Apr/2016:22:49:09 -0400] "GET /spring2016/keep_alive HTTP/1.1" 200 1012 "http://15418.courses.cs.cmu.edu/spring2016/article/15" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_3) AppleWebKit/537.36 (KHTML, like Gecko) Chr
```

The log of page views gets quite large...

Assume log.txt is a large file, stored in a distributed file system, like HDFS
Below: cluster of 4 nodes,
each node with a 1 TB disk
Contents of log.txt are distributed evenly in blocks across the cluster



Example query:

"What type of mobile phone are all the visitors using?"

Using MapReduce

```
// called once per line in input file by runtime
// input: string (line of input file)
// output: adds (user_agent, 1) entry to list
void mapper(string line, multimap<string,string>& results) {
   string user_agent = parse_requester_user_agent(line);
   if (is_mobile_client(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://15418log.txt");

// stores output
Writer output("hdfs://…");

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

The code left computes the count of page views by each type of mobile phone.