


mrjob

Jimmy Retzlaff
Yelp


What's the Problem?

- Yelp produces ~100GB of logs per day
- Many features rely on log analysis
- Computers are cheap, but leveraging many of them for a single task is hard



Real people. Real reviews.®

[Welcome](#)
[About Me](#)
[Write a Review](#)
[Find Reviews](#)
[Invite Friends](#)
[Messaging](#)
[Talk](#)
[Events](#)
[Member Search](#)



100% ALL NATURAL. BECAUSE NOT ALL MILK CHOCOLATE IS CREATED EQUAL.


[Indulge >](#)

Pier 14



14 reviews
[Rating Details](#)

Category: [Parks](#)
[Edit](#)


The Embarcadero and Mission Street
San Francisco, CA 94105

Good for Kids: Yes
[Edit Business Info](#)

[Send to Friend](#)
[Bookmark](#)
[Send to Phone](#)
[Write a Review](#)
[Print](#)




Gray Line / San Francisco Sights...
San Francisco's Premiere Sightseeing Company!
<http://SANFRANCISCO-SIGHTSEEING.COM>



LaLanne Fitness CrossFit
CrossFit in San Francisco Serious Workout. Small Classes.
<http://lallanefitness.reachlocal.com/?sclid=1830115>


Ads by CityGrid



[Add Photos](#)

[First to Review](#)

Toro E.



[View Larger Map/Directions >](#)


Browse Nearby:
[Restaurants](#) | [Nightlife](#) | [Shopping](#) | [Movies](#) | [All](#)

People Who Viewed This Also Viewed...




Fort Point


137 reviews
San Francisco, CA




Pier 7

10 reviews
Neighborhood: Embarcadero



Mission Creek Park

20 reviews
Neighborhood: SOMA




Pier 37

1 review
Neighborhood: Financial District

14 reviews for Pier 14

Sort by: [Yelp Sort](#) | [Date](#) | [Rating](#) | [Elites](#) | [Facebook Friends](#)

All Reviews



Elite '10

67
477
Kylie L.
San Francisco, CA


8/6/2010
5 photos

Why is pier 14 <http://www.yelp.com/bl...> next to pier 2 & not next to pier 15?

The view is AMAZING, especially on a sunny day. Great views of the Bay Bridge <http://www.yelp.com/bl...>, Port of SF, Coit Tower <http://www.yelp.com/bl...>, kayakers, boaters, sailors <http://www.yelp.com/bl...>, & people fishing <http://www.yelp.com/bl...>. There are individual rotating seats along the pier, which I think is really cool.

pier 14 is cool!

Ads by CityGrid

ous Workout. Small Classes.
com/?scid=1830115

Search Reviews


er 15?

Bay Bridge
bi..., kayakers, boaters,
n/bi.... There are individual


View Larger Map/Directions »

Browse Nearby:
Restaurants | Nightlife | Shopping | Movies |
All


People Who Viewed This Also Viewed...




Fort Point
★★★★★ 137 reviews
San Francisco, CA



Pier 7
★★★★★ 10 reviews
Neighborhood: Embarcadero



Mission Creek Park
★★★★★ 20 reviews
Neighborhood: SOMA



Pier 37
★★★★★ 1 review
Neighborhood: Financial District

how do we do this? walk through our logs and collect stats about pairs of businesses viewed in the same browsing session – but it's 100GB!



Go, Dog. Go!

go
dog
go

go \longrightarrow counter() \longrightarrow (g, l), (o, l)
dog \longrightarrow counter() \longrightarrow (d, l), (o, l), (g, l)
go \longrightarrow counter() \longrightarrow (g, l), (o, l)

$(g, l), (o, l)$
 $(d, l), (o, l), (g, l)$
 $(g, l), (o, l)$ \longrightarrow sort \longrightarrow (d, l)
 (g, l)
 (g, l)
 (g, l)
 (o, l)
 (o, l)
 (o, l)

(d, l)

(g, l)

(g, l)

(g, l)

(o, l)

(o, l)

(o, l)

(d, 1)	→	summarizer(d, [1])	→	(d, 1)
<hr/>				
(g, 1)				
(g, 1)	→	summarizer(g, [1, 1, 1])	→	(g, 3)
(g, 1)				
<hr/>				
(o, 1)				
(o, 1)	→	summarizer(o, [1, 1, 1])	→	(o, 3)
(o, 1)				

hey, this could be run on multiple computers!

go \longrightarrow counter() \longrightarrow (g, l), (o, l)
dog \longrightarrow counter() \longrightarrow (d, l), (o, l), (g, l)
go \longrightarrow counter() \longrightarrow (g, l), (o, l)

go	→	counter()	→	(g, l), (o, l)
<hr/>				
dog	→	counter()	→	(d, l), (o, l), (g, l)
<hr/>				
go	→	counter()	→	(g, l), (o, l)

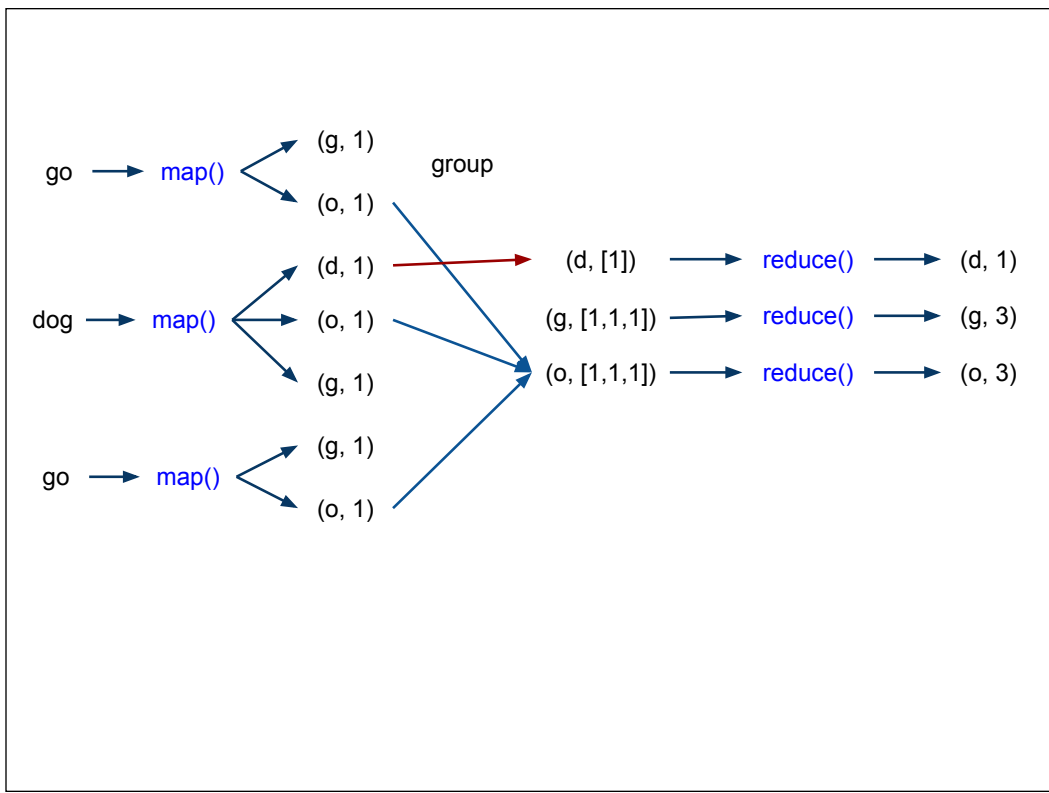
how about a different name? mapper

go	→	mapper()	→	(g, l), (o, l)
<hr/>				
dog	→	mapper()	→	(d, l), (o, l), (g, l)
<hr/>				
go	→	mapper()	→	(g, l), (o, l)

(d, 1)	→	summarizer(d, [1])	→	(d, 1)
<hr/>				
(g, 1)				
(g, 1)	→	summarizer(g, [1, 1, 1])	→	(g, 3)
(g, 1)				
<hr/>				
(o, 1)				
(o, 1)	→	summarizer(o, [1, 1, 1])	→	(o, 3)
(o, 1)				

how about a different name? reducer

(d, 1)	→	reducer(d, [1])	→	(d, 1)
<hr/>				
(g, 1)				
(g, 1)	→	reducer(g, [1, 1, 1])	→	(g, 3)
(g, 1)				
<hr/>				
(o, 1)				
(o, 1)	→	reducer(o, [1, 1, 1])	→	(o, 3)
(o, 1)				



Yelp's mrjob

- sign up for Amazon Elastic MapReduce
- write a dozen lines of Python
- run on many Amazon computers

mrjob takes care of the plumbing – it launches the Amazon servers for you, let's you install your code on them, uploads your data, kicks off your job, monitors it, brings the results back, and shuts the servers down

```
from mrjob.job import MRJob

class MRCharacterCount(MRJob):
    def mapper(self, _, text):
        for c in text:
            yield c, 1

    def reducer(self, c, counts):
        yield c, sum(counts)

if __name__ == '__main__':
    MRCharacterCount.run()
```

```
from mrjob.job import MRJob

class MRWordCount(MRJob):
    def mapper(self, _, text):
        for word in text.split():
            yield word, 1

    def reducer(self, word, counts):
        yield word, sum(counts)

if __name__ == '__main__':
    MRWordCount.run()
```

The Great American Cheese Collection

Category: [Farmers Market](#)

Neighborhood: [Brighton Park](#)

Sponsored Result

★ ★ ★ ★ ★

9 reviews

4727 S Talman Ave
Chicago, IL 60632
(773) 779-5055

🍷 50% off selected cheeses

Free Cheese Tasting EVERY Saturday. 10 AM-1 PM at our warehouse. 4727 S. Talman.... [read more »](#)

click through rate = ads clicked / ads shown

show ad_event_log in BBEdit

Exercise 1: Run an mrjob!

http://snipurl.com/mr_word_count

\$ python mr_word_count.py mr_word_count.py

```
from mrjob.job import MRJob

class MRWordCount(MRJob):
    def mapper(self, _, text):
        for word in text.split():
            yield word, 1

    def reducer(self, word, counts):
        yield word, sum(counts)

if __name__ == '__main__':
    MRWordCount.run()
```

Exercise 2: Write an mrjob!

Determine the number of times every two characters appear in order in the same word in some text.

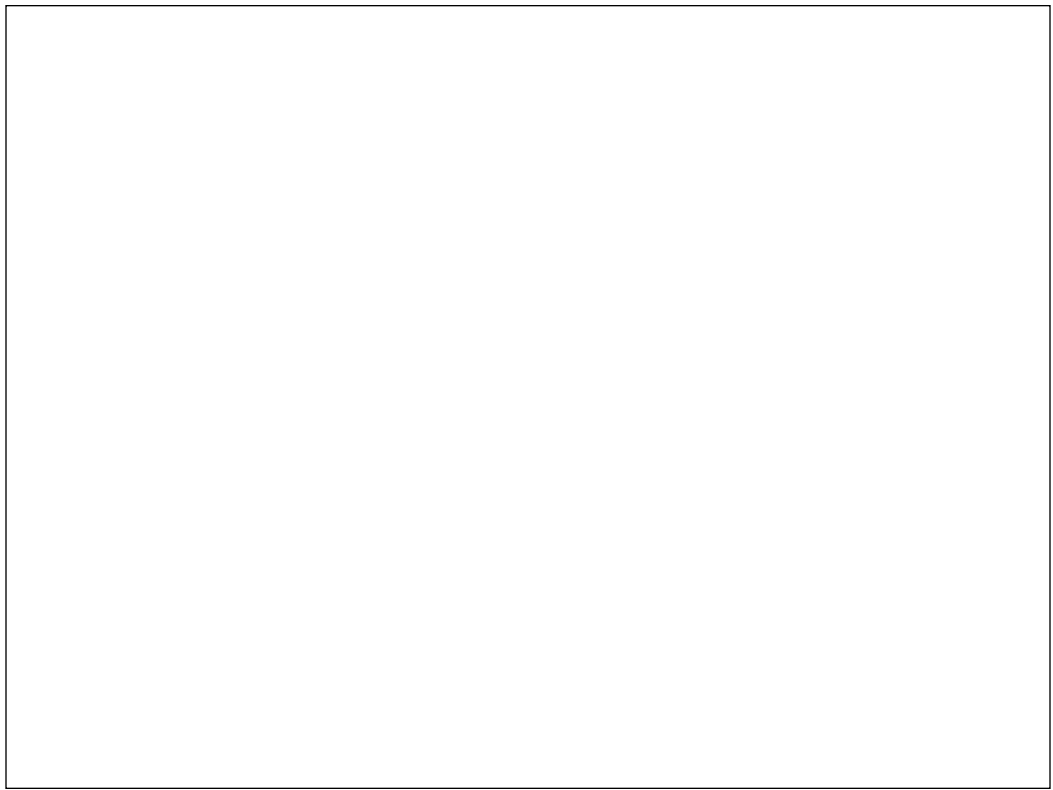
Example: go dog go

g,o 2

d,o 1

o,g 1

Extra credit: After doing the first part, modify it so that order does not matter.... g,o 3; d,o 1



Where can I get cool data to play with?

Where can I get cool data to play with?

Yelp is hiring!

[yelp.com/jobs](https://www.yelp.com/jobs)