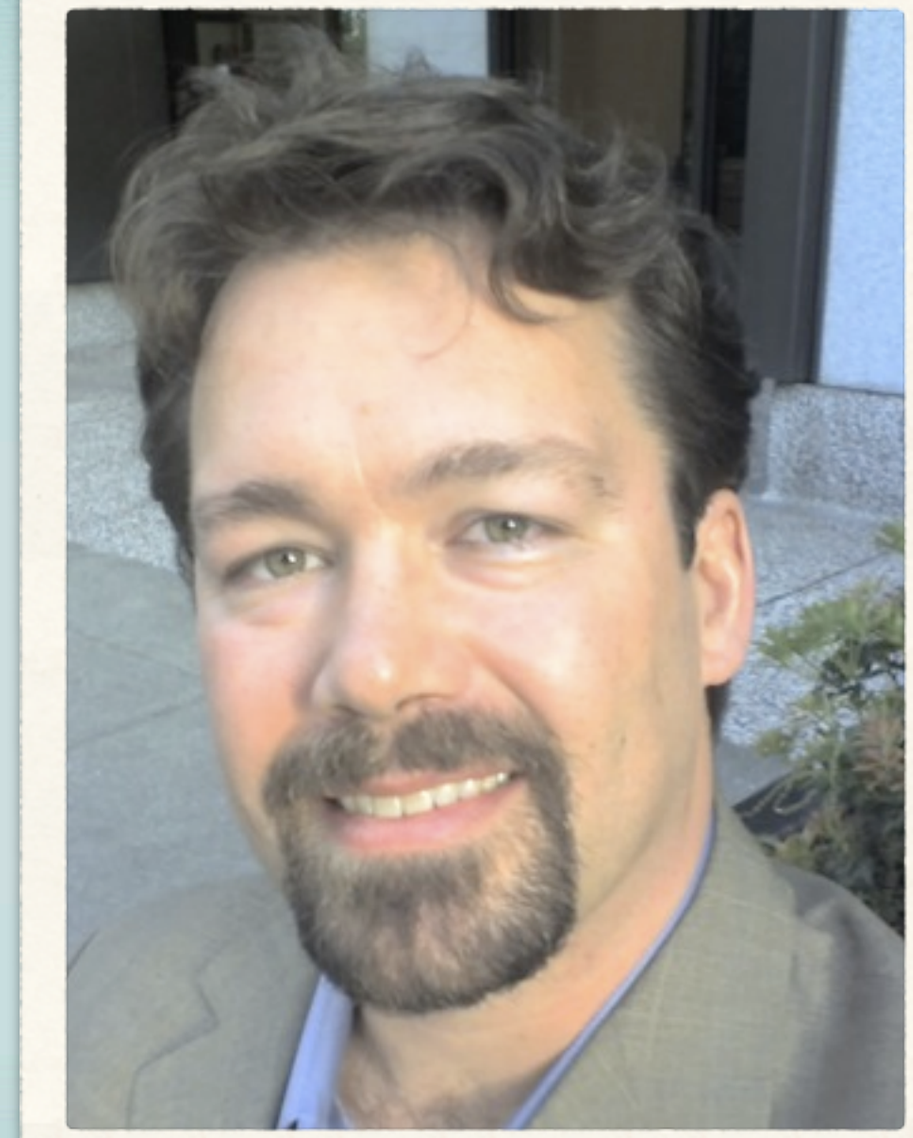
ENRON EMAIL & COUCHDB

Adventures in Lisp (Erlang), Ruby & Javascript

Tim Dysinger

Positive Inertia,
Principle Consultant

Sonian Networks,
VP of Engineering



What's the experiment?

* Test: Use CouchDB end-to-end in a real-world situation

* Dependency: A large pile of semi-structured documents
[Enron Email]

```ruby
%w(tmail find restclient json).each { |l| require l }
db = "http://localhost:5984/enron"
RestClient.put(db, '') rescue nil
Find.find('maildir') do |path|
  next if FileTest.directory?(path)
  begin
    msg = TMail::Mail.parse(IO.read(path))
    attrs = msg.header.merge('to' => msg.to_addrs,
                             'cc' => msg.cc_addrs,
                             'bcc' => msg.bcc_addrs,
                             'body' => msg.body).reject {|k,v| v.to_s.empty?}
    RestClient.post(db, attrs.to_json, :content_type => 'application/json')
  rescue Interrupt
    exit(1)
  rescue Exception => ex
    puts "#{path} #{ex.inspect}"
  end
end
```

Upload Email into CouchDB With Ruby

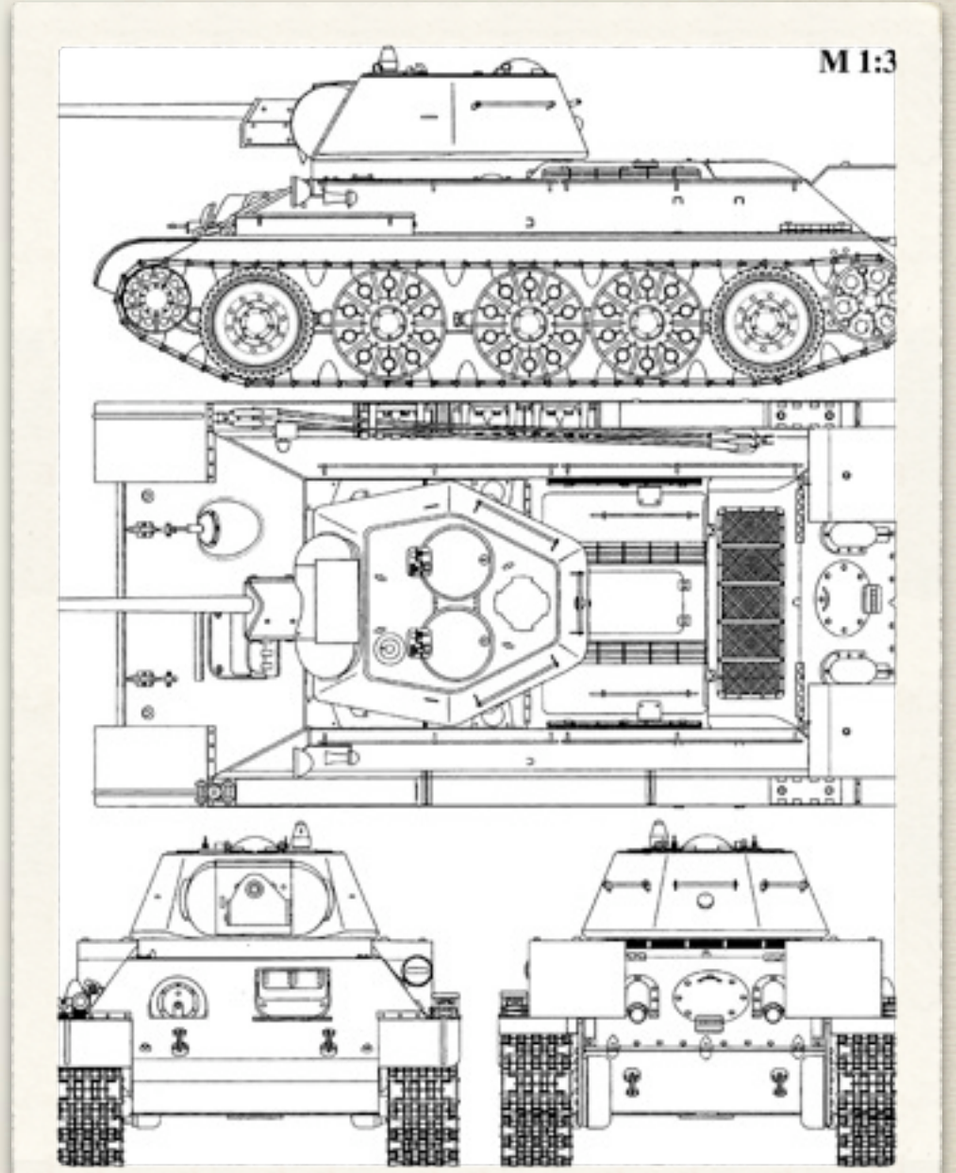
Field	Value
_id	"0021288bfada8b08f9803509a267faae"
_rev	"1-410990586"
body	"John, Please disregard the prior version and use this one. Thanks.\n"
content-transfer-encoding	"7bit"
content-type	"text/plain; charset=us-ascii"
date	"Mon, 30 Apr 2001 09:01:00 -1000"
from	"gerald.nemec@enron.com"
message-id	"<22219331.1075842910570.JavaMail.evans@thyme>"
mime-version	"1.0"
subject	"Revised CSA"
to	0 "john.kiani@enron.com"

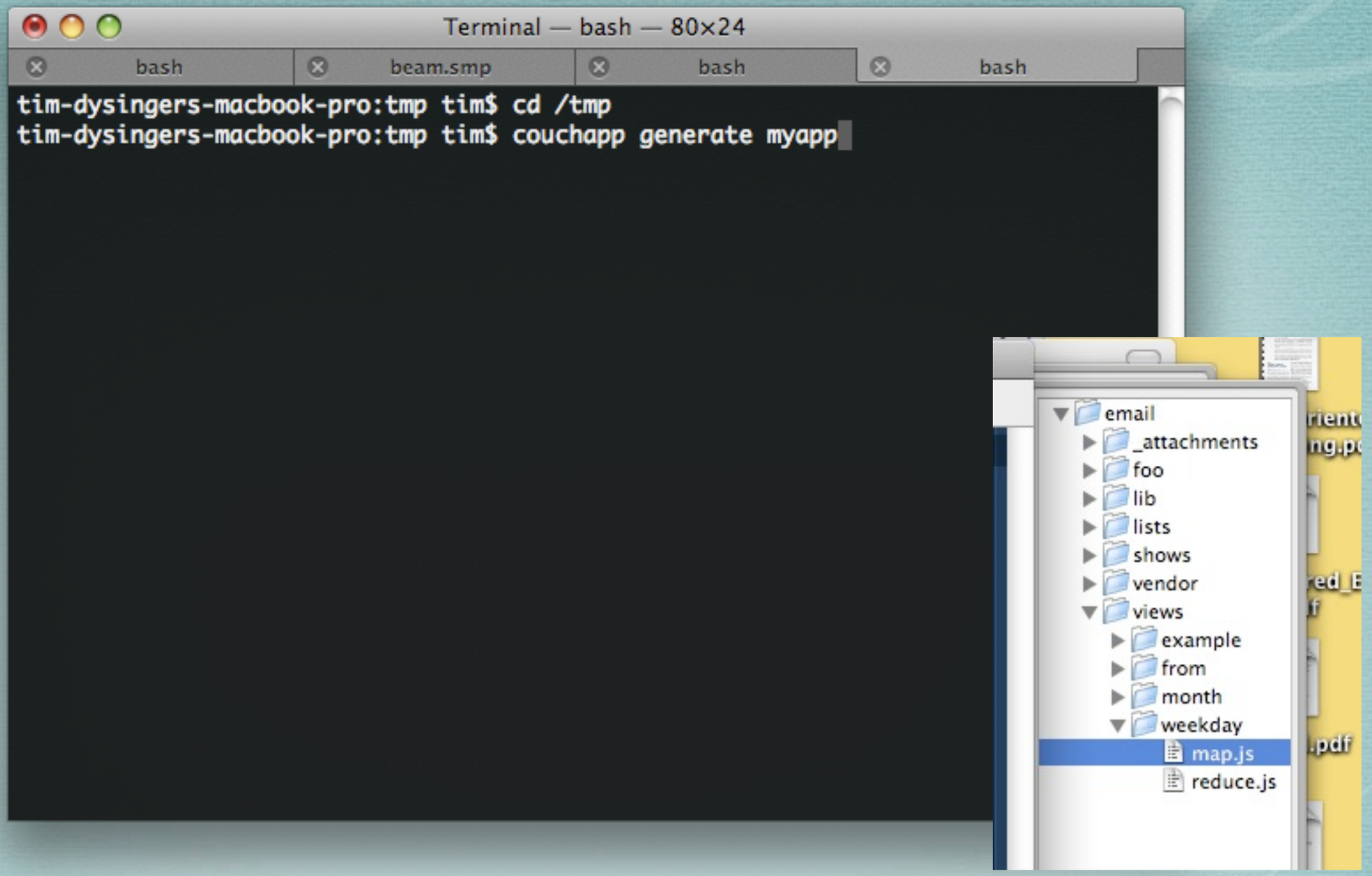
Example Email

As seen relaxing on the futon

What now? (the plan)

- Set up our CouchApp
- Create some map/reduce views
- Use our views to extract info and create reports





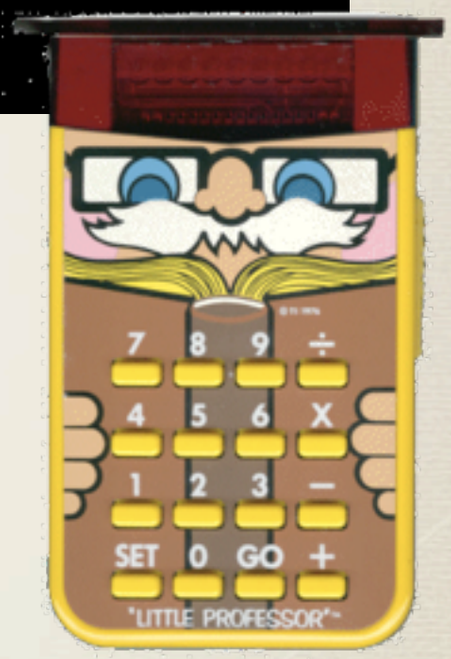
Creating a CouchApp

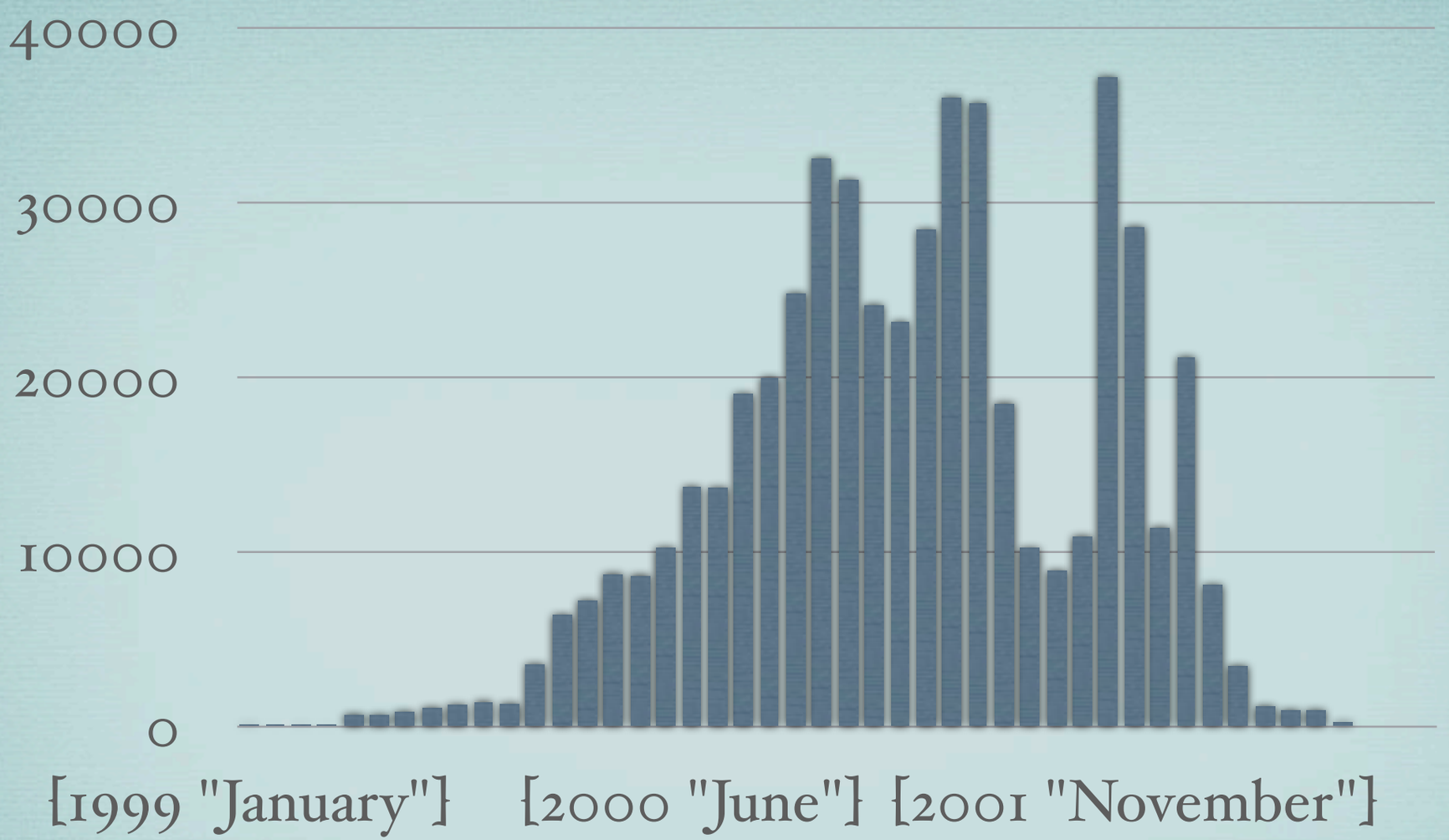
Our first view

```
// !code lib/helpers/date.js

function(doc) {
  try {
    emit(Date.parse(doc.date).getDayName(),1);
  } catch (error) {
    log([doc._id,doc.date]);
    emit("Unprocessable",1);
  }
}
```

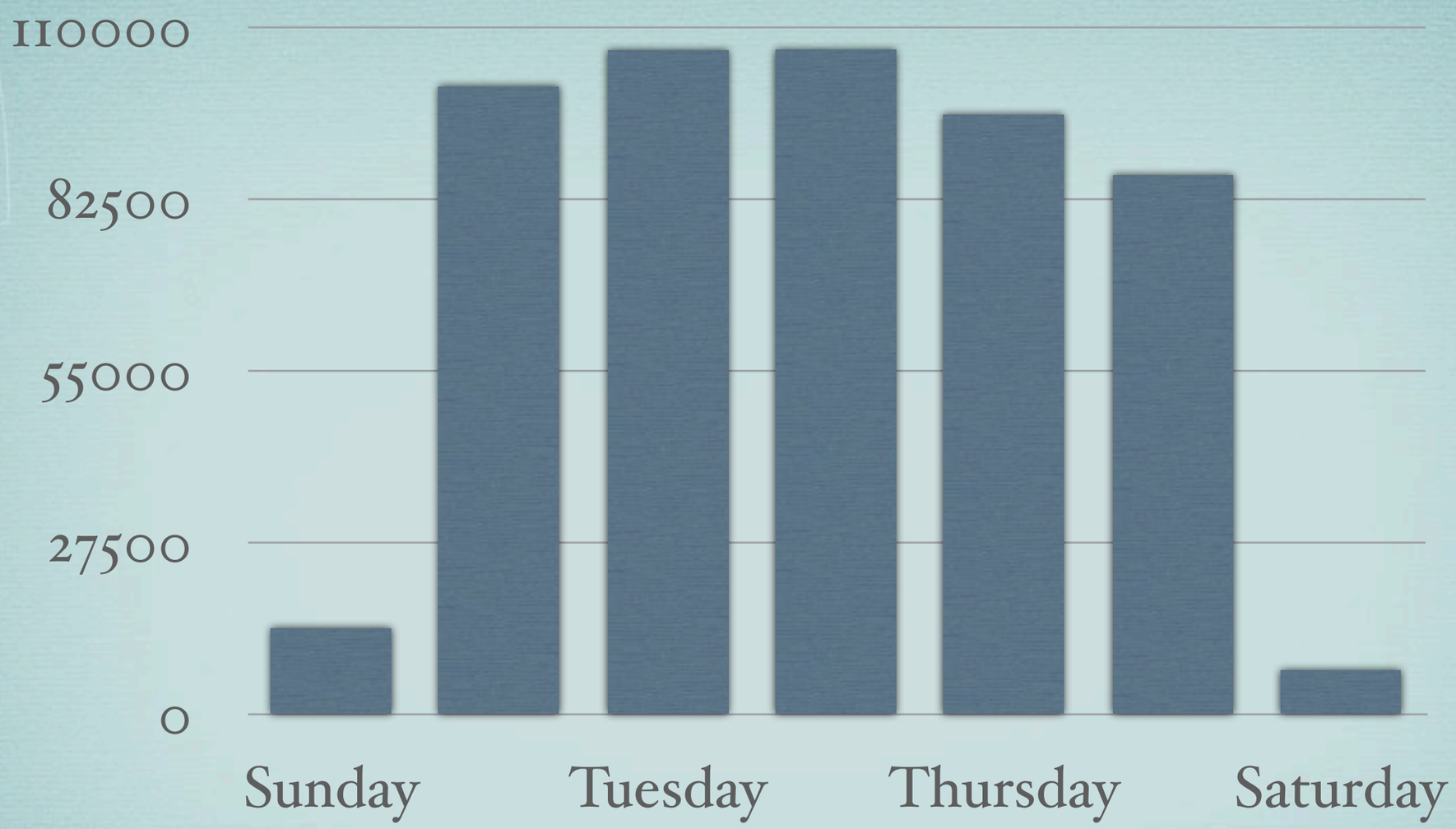
```
function(key, values) {
  return sum(values);
}
```





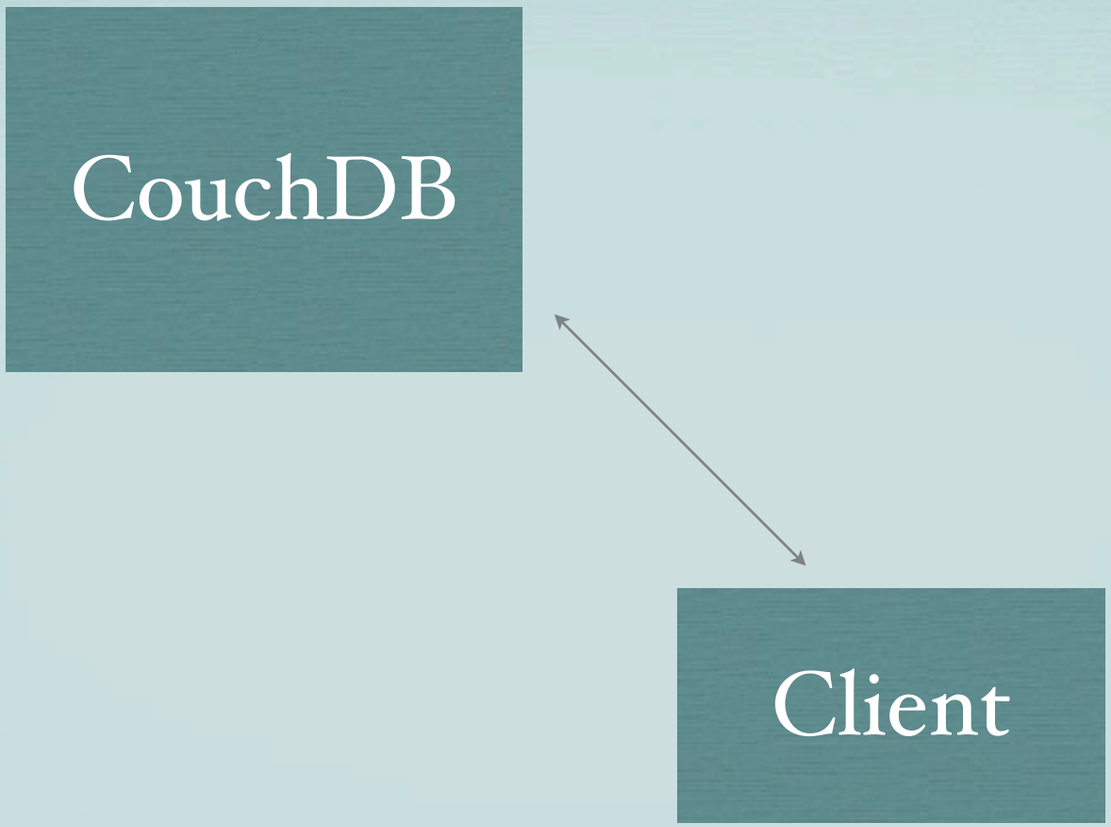
Map/Reduce Sum

Emails by Month from 1999-2002



Map/Reduce Sum

Emails by Day of Week from 1999-2002

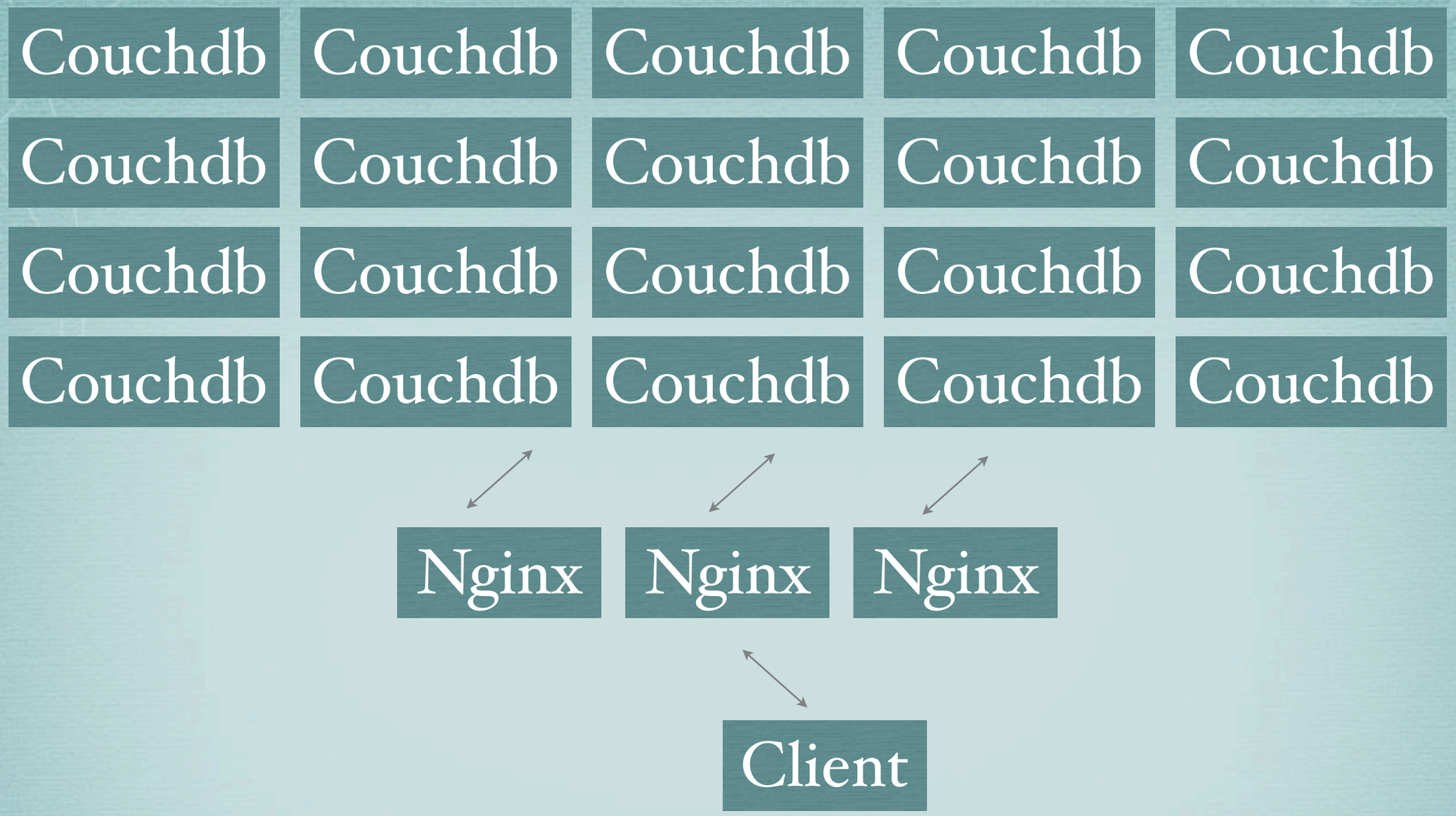


A single node (100-ish/s)

CouchDB

Uploader

A single node (faster^10)



Multiple nodes

Hashing to CouchDB

Issues

* Compression?
* Distribution?