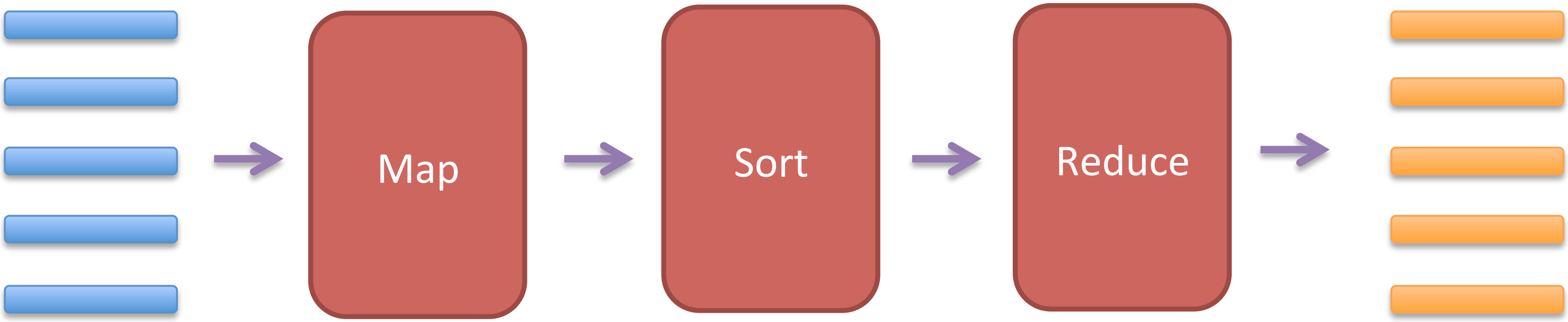
Can we make 🟥 run anywhere?

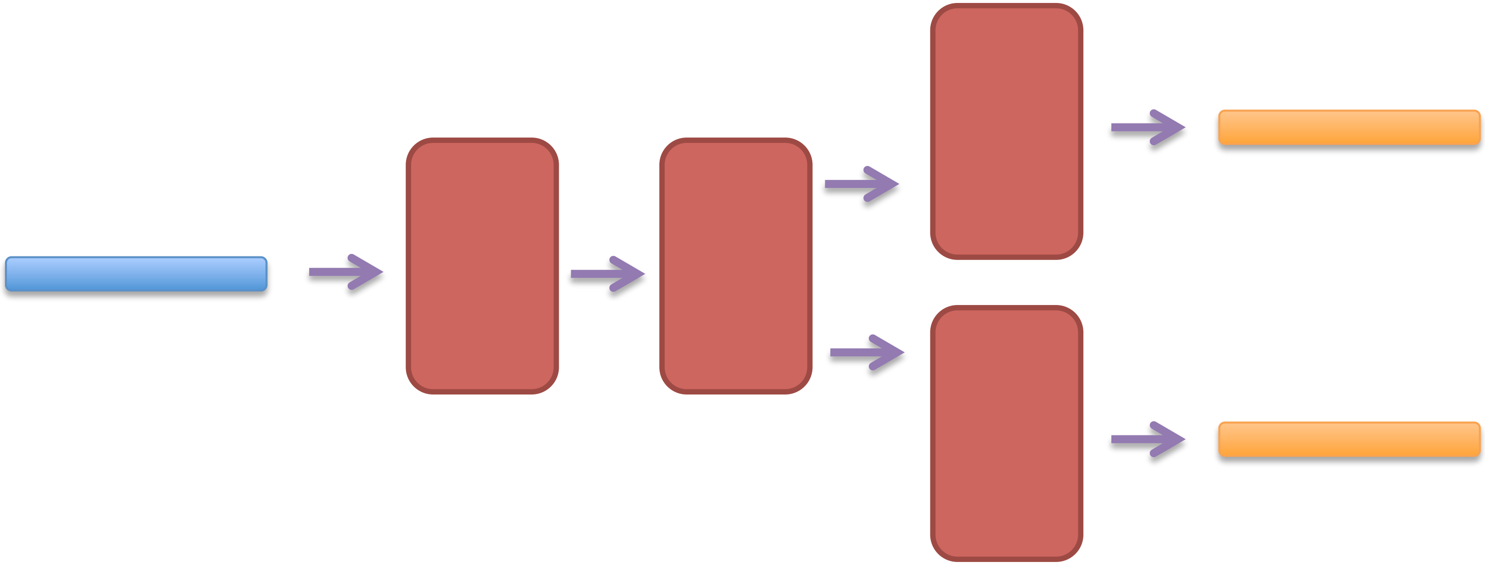
Dhruv Bansal

infochimps

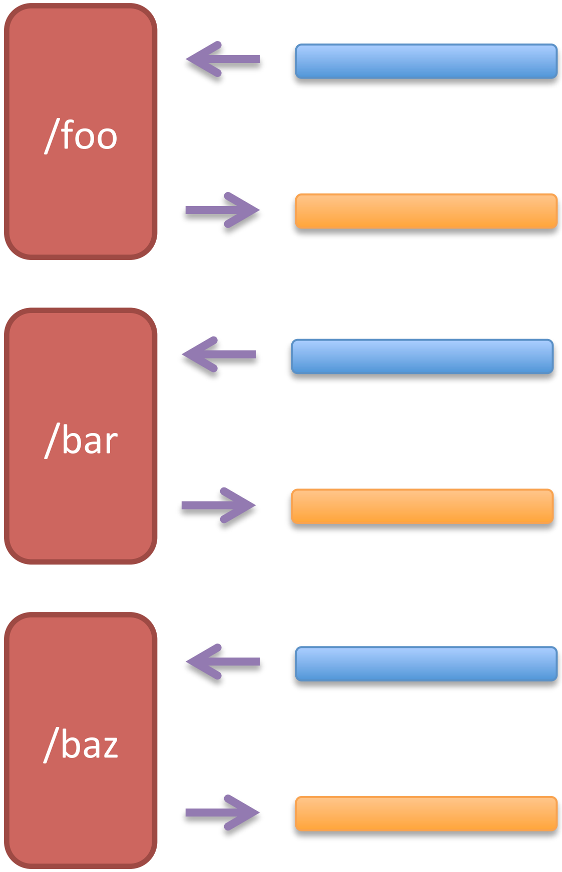
What is a Hadoop job?



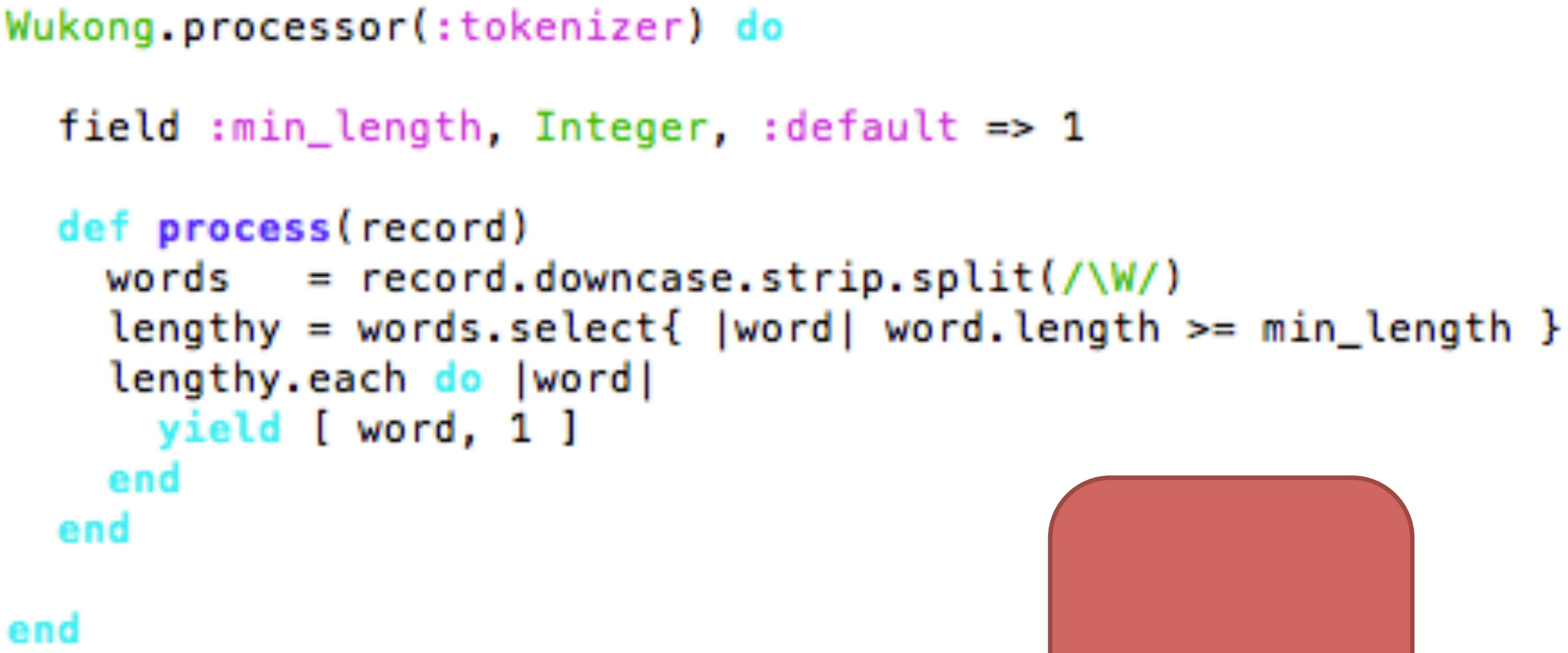
What is Storm?



What is a SOA?

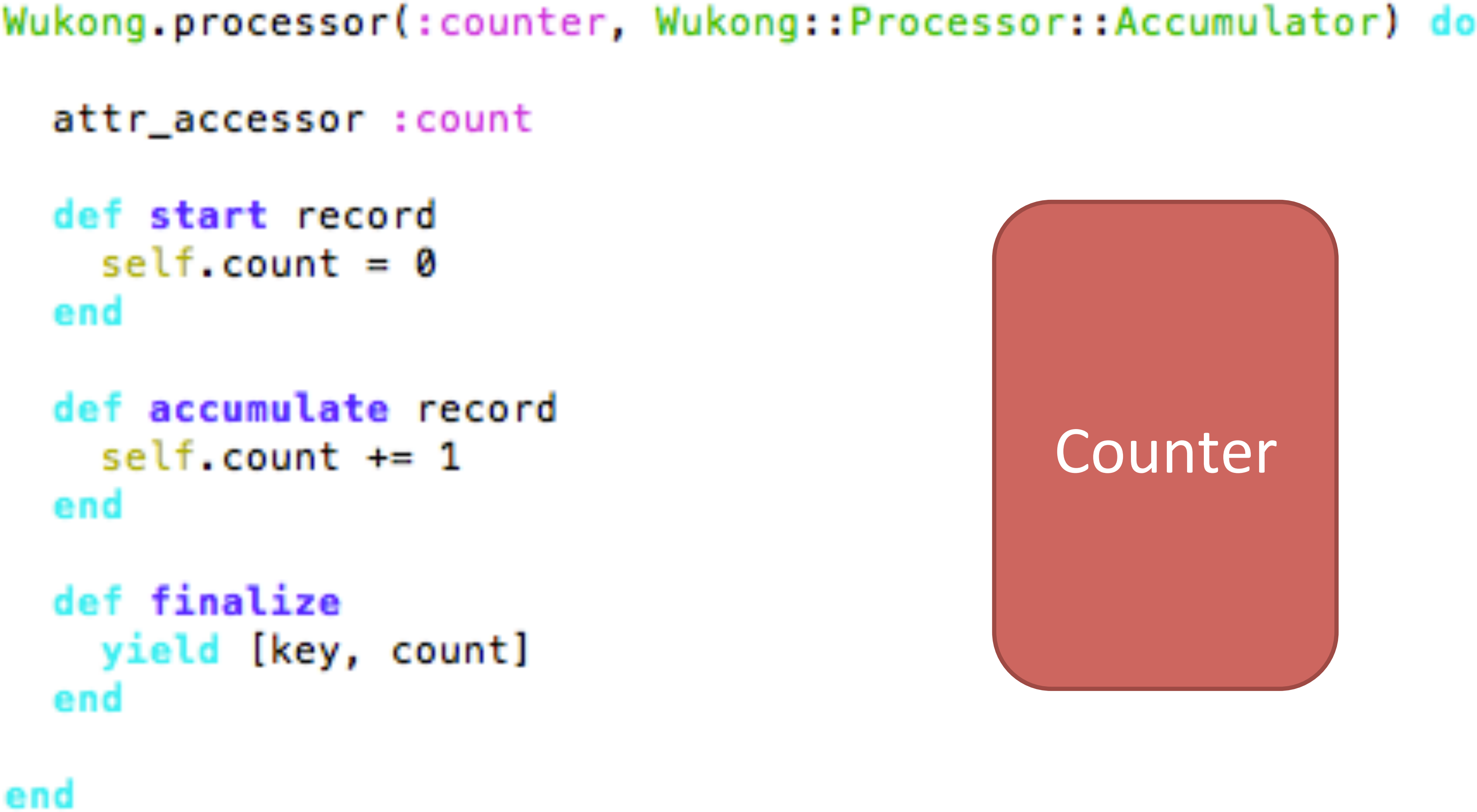


Can we make 🟥 run anywhere?



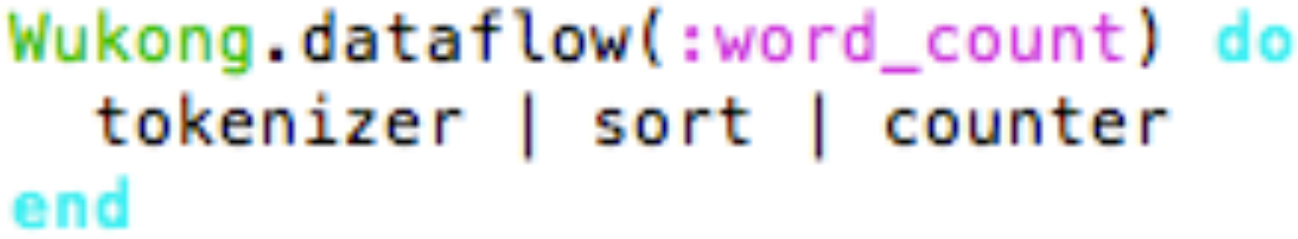
```
Wukong.processor(:tokenizer) do

  field :min_length, Integer, :default => 1

  def process(record)
    words   = record.downcase.strip.split(/\W/)
    lengthy = words.select{ |word| word.length >= min_length }
    lengthy.each do |word|
      yield [ word, 1 ]
    end
  end

end
```

Tokenizer

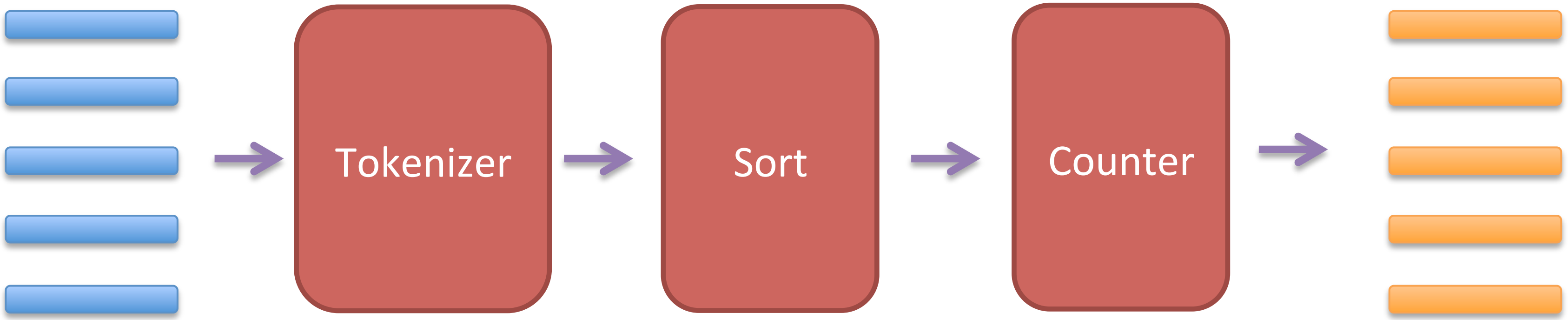


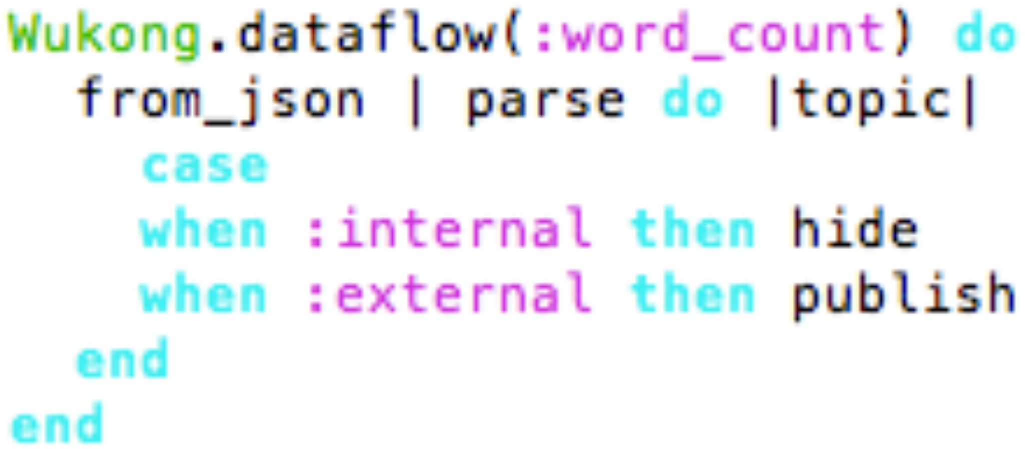
```
Wukong.processor(:counter, Wukong::Processor::Accumulator) do

  attr_accessor :count

  def start record
    self.count = 0
  end

  def accumulate record
    self.count += 1
  end

  def finalize
    yield [key, count]
  end

end
```

Counter

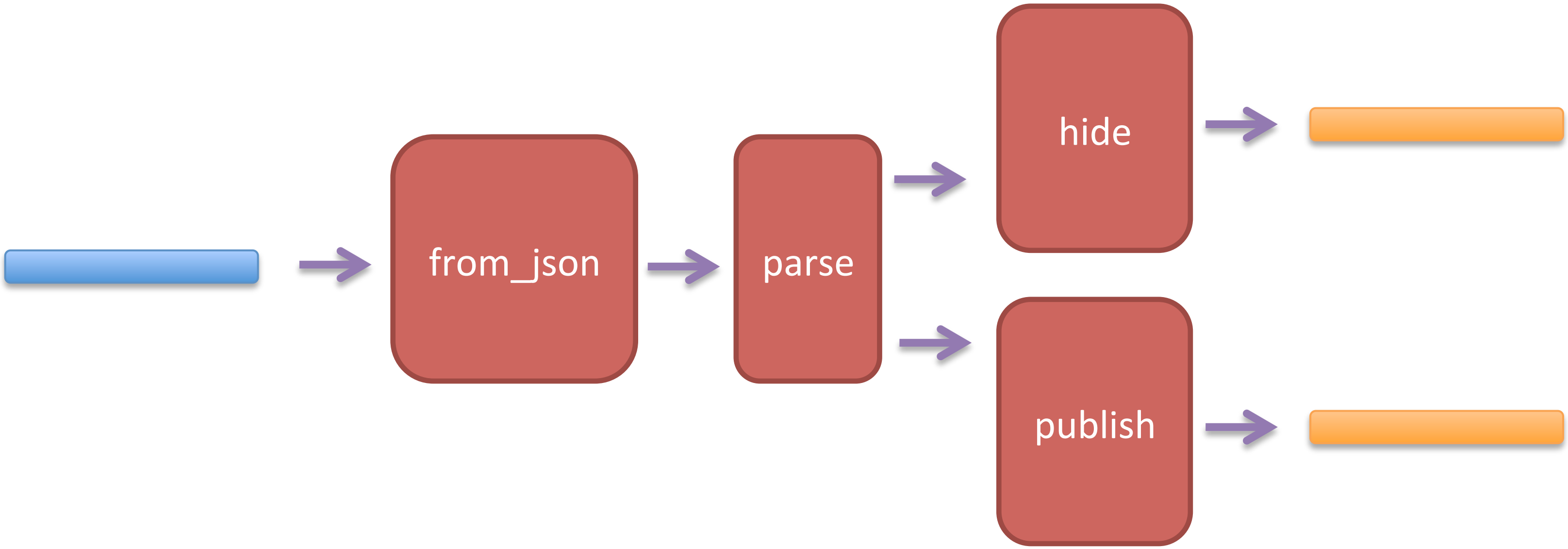


```
Wukong.dataflow(:word_count) do
  tokenizer | sort | counter
end
```

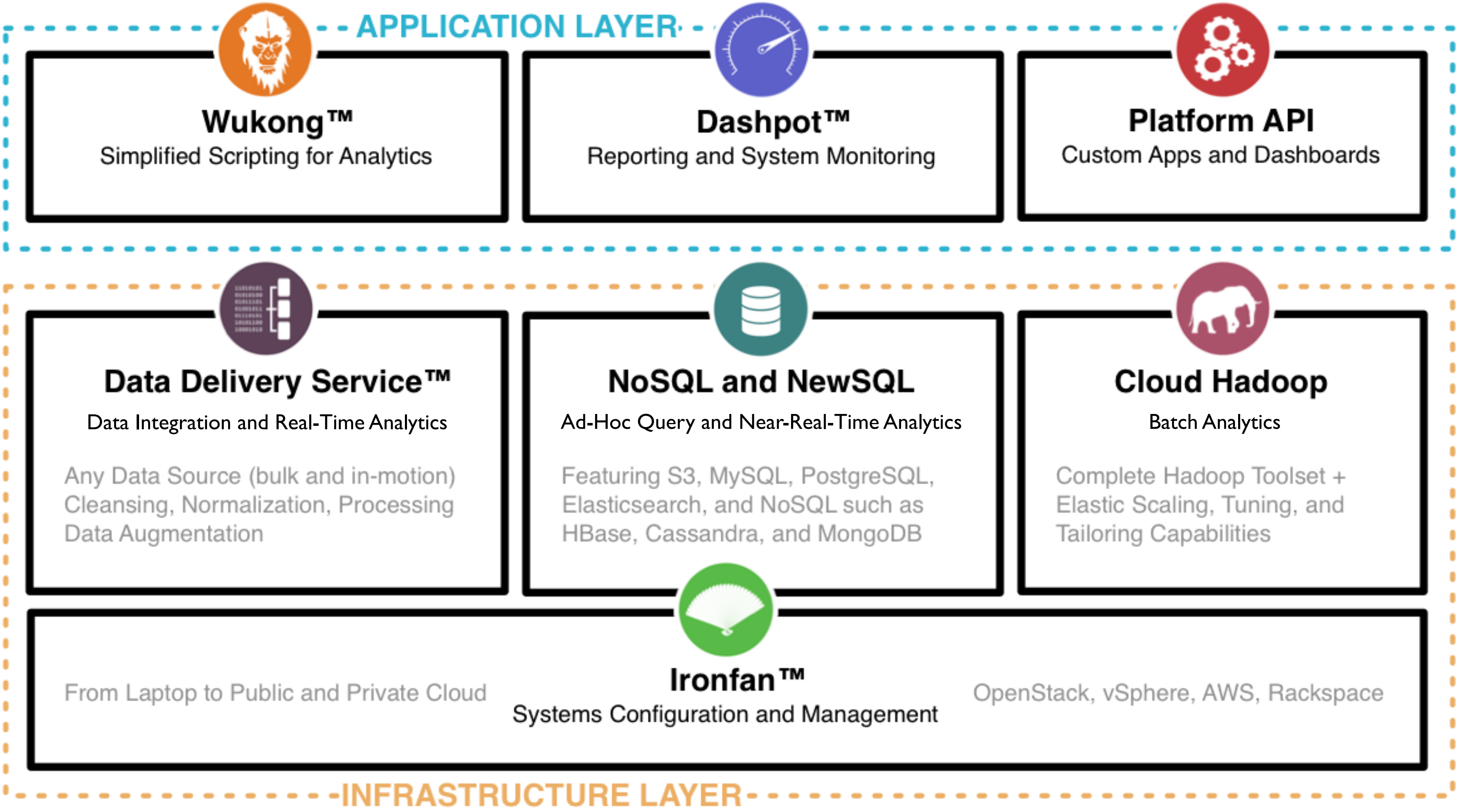




```
Wukong.dataflow(:word_count) do
  from_json | parse do |topic|
    case
    when :internal then hide
    when :external then publish
  end
end
```



Infochimps Platform Components



APPLICATION LAYER

Wukong™
Simplified Scripting for Analytics

Dashpot™
Reporting and System Monitoring

Platform API
Custom Apps and Dashboards

Data Delivery Service™
Data Integration and Real-Time Analytics

Any Data Source (bulk and in-motion)
Cleansing, Normalization, Processing
Data Augmentation

NoSQL and NewSQL
Ad-Hoc Query and Near-Real-Time Analytics

Featuring S3, MySQL, PostgreSQL, Elasticsearch, and NoSQL such as HBase, Cassandra, and MongoDB

Cloud Hadoop
Batch Analytics

Complete Hadoop Toolset + Elastic Scaling, Tuning, and Tailoring Capabilities

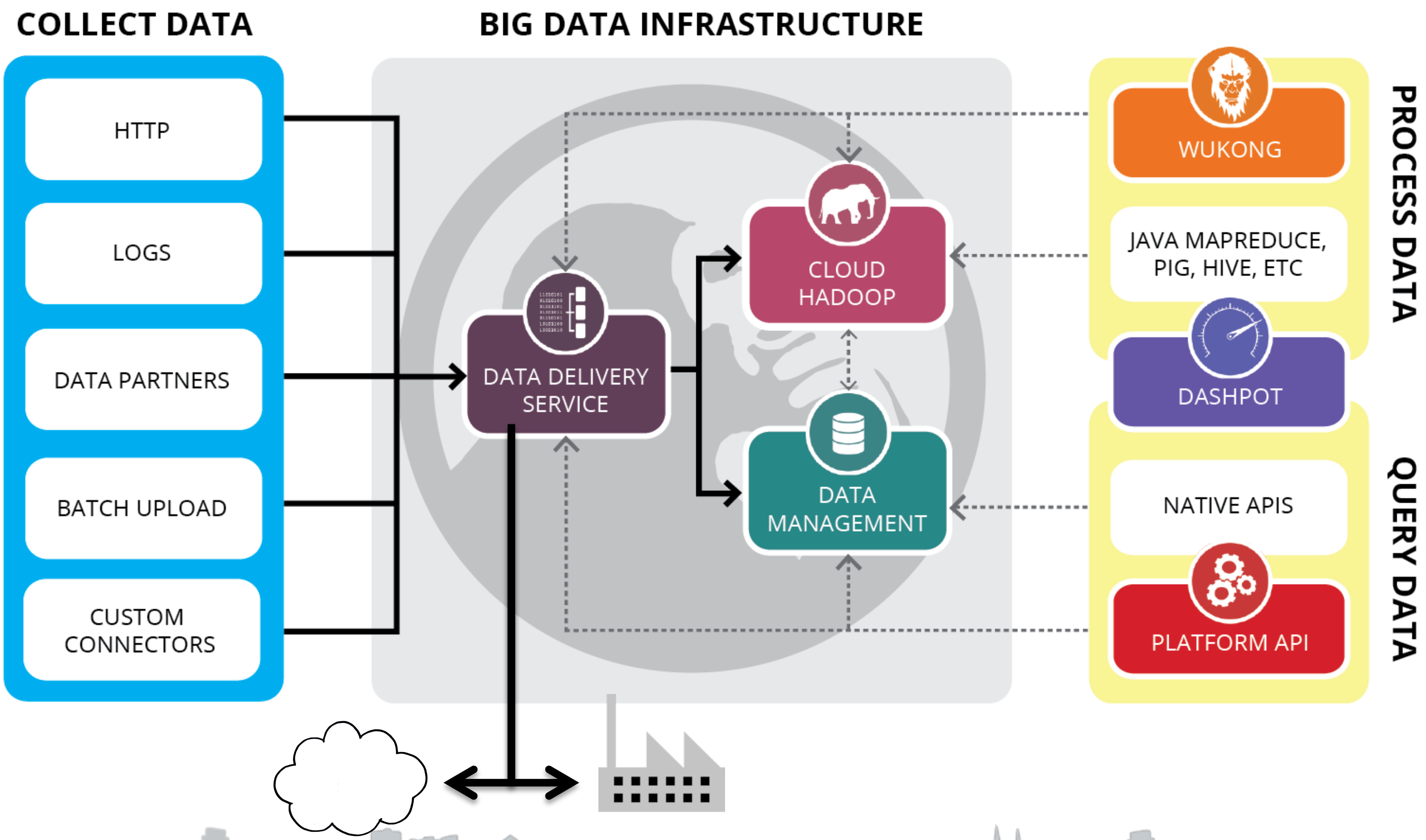
From Laptop to Public and Private Cloud

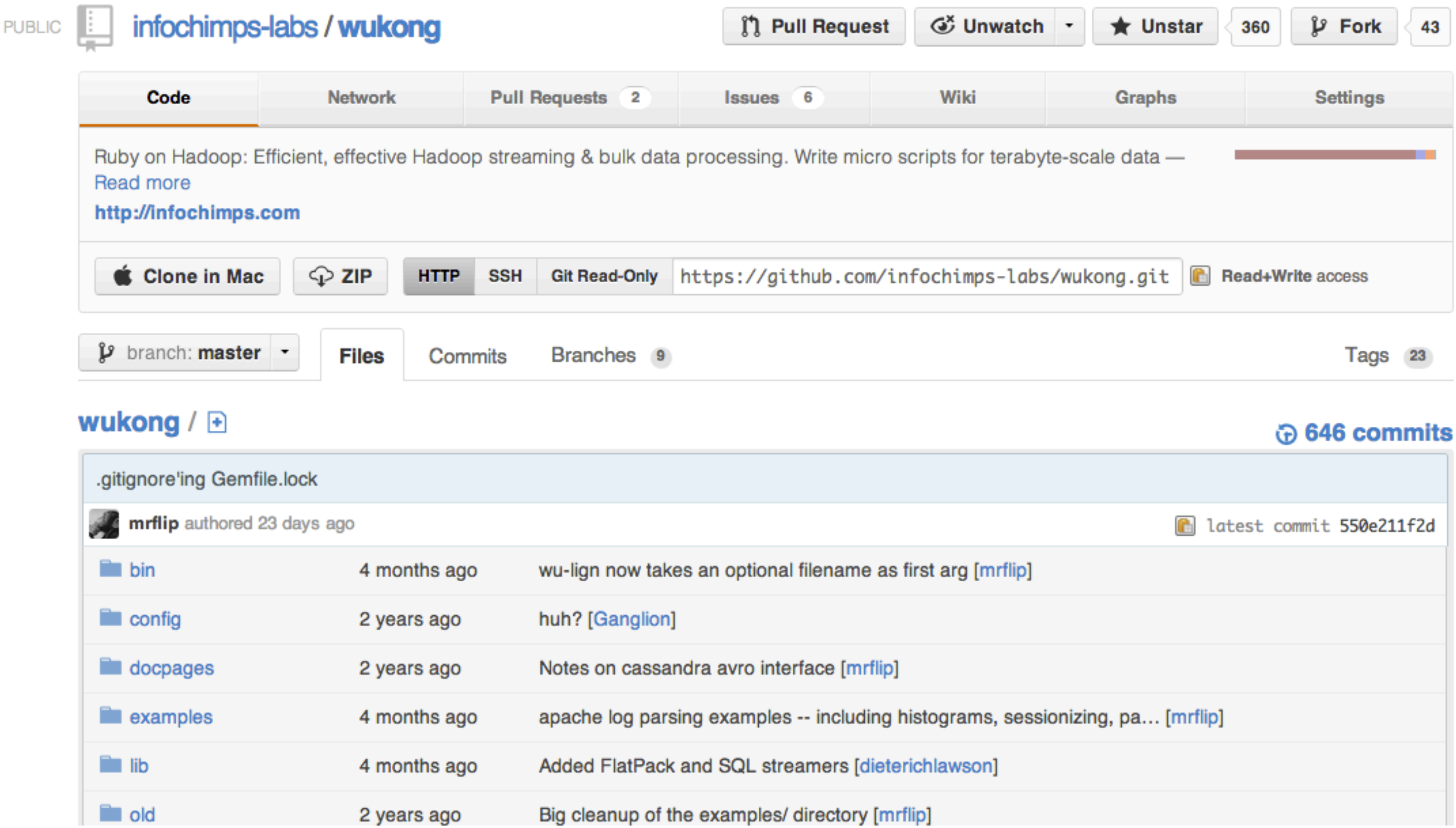
Ironfan™
Systems Configuration and Management

OpenStack, vSphere, AWS, Rackspace

INFRASTRUCTURE LAYER

Infochimps Platform Architecture





Ruby on Hadoop: Efficient, effective Hadoop streaming & bulk data processing. Write micro scripts for terabyte-scale data — Read more

http://infochimps.com

https://github.com/infochimps-labs/wukong.git

wukong /

646 commits

.gitignore'ing Gemfile.lock

mrflip authored 23 days ago — latest commit 550e211f2d

bin	4 months ago	wu-lign now takes an optional filename as first arg [mrflip]
config	2 years ago	huh? [Ganglion]
docpages	2 years ago	Notes on cassandra avro interface [mrflip]
examples	4 months ago	apache log parsing examples -- including histograms, sessionizing, pa… [mrflip]
lib	4 months ago	Added FlatPack and SQL streamers [dieterichlawson]
old	2 years ago	Big cleanup of the examples/ directory [mrflip]

https://github.com/infochimps-labs