bigdata, with 0% java

김태웅
kdog@exabytes.kr

@ python / mac /go 강제 에반젤리스트
@ S*/K* 등 통신사 쪽 경험이 多

목차

what is big data?

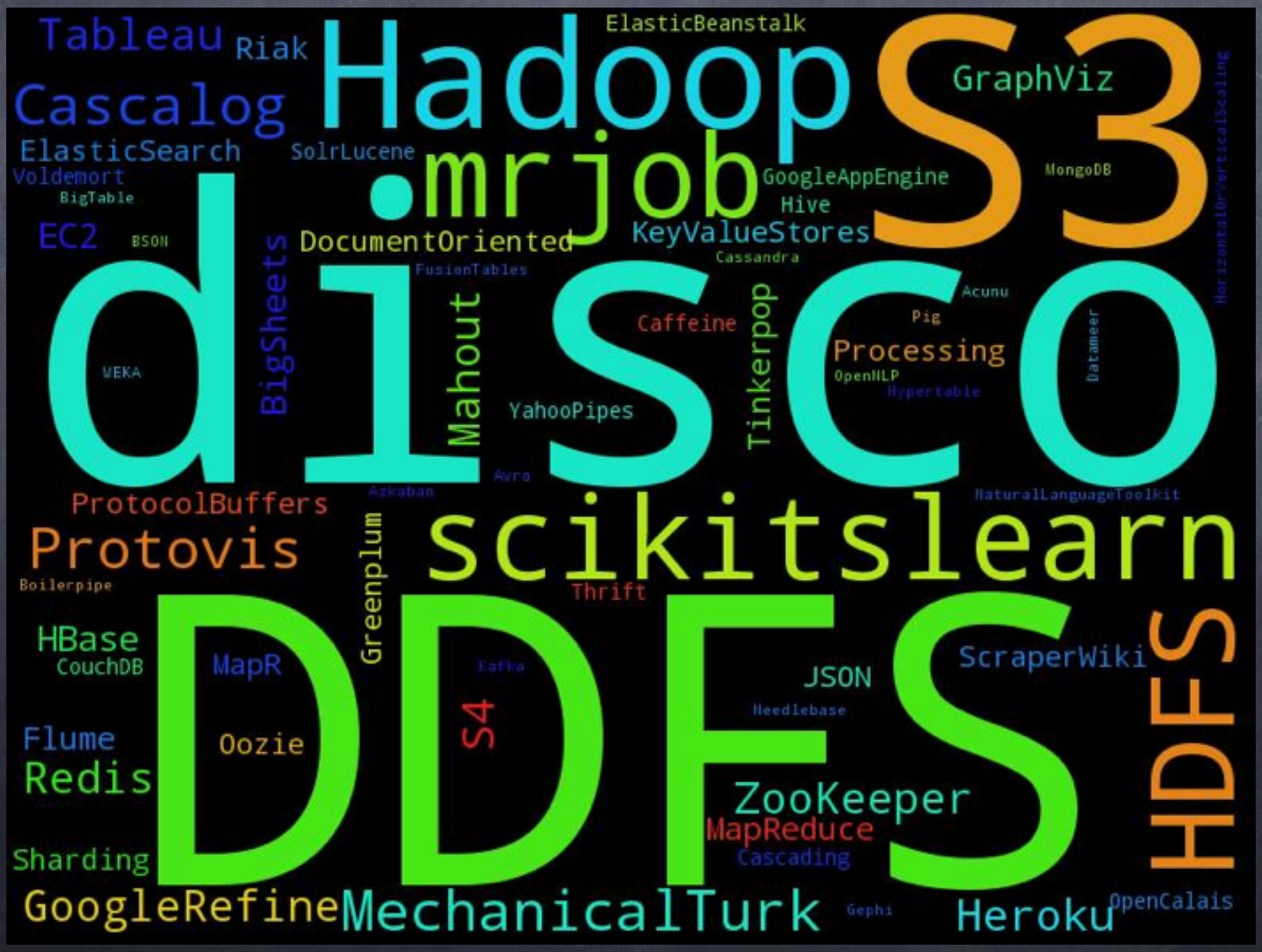
@ Volume(규모)
@ Variety(다양성)
@ Velocity(속도)
@ Value(가치)

마케팅 용어인가요?

@ 10년전 떡밥 : web 2.0
@ "기술적" 관점에서의 빅 데이터
 - map reduce
@ 데이터 기반 의사결정을 할 수 있게 해주는 보조도구
 - microsoft excel?

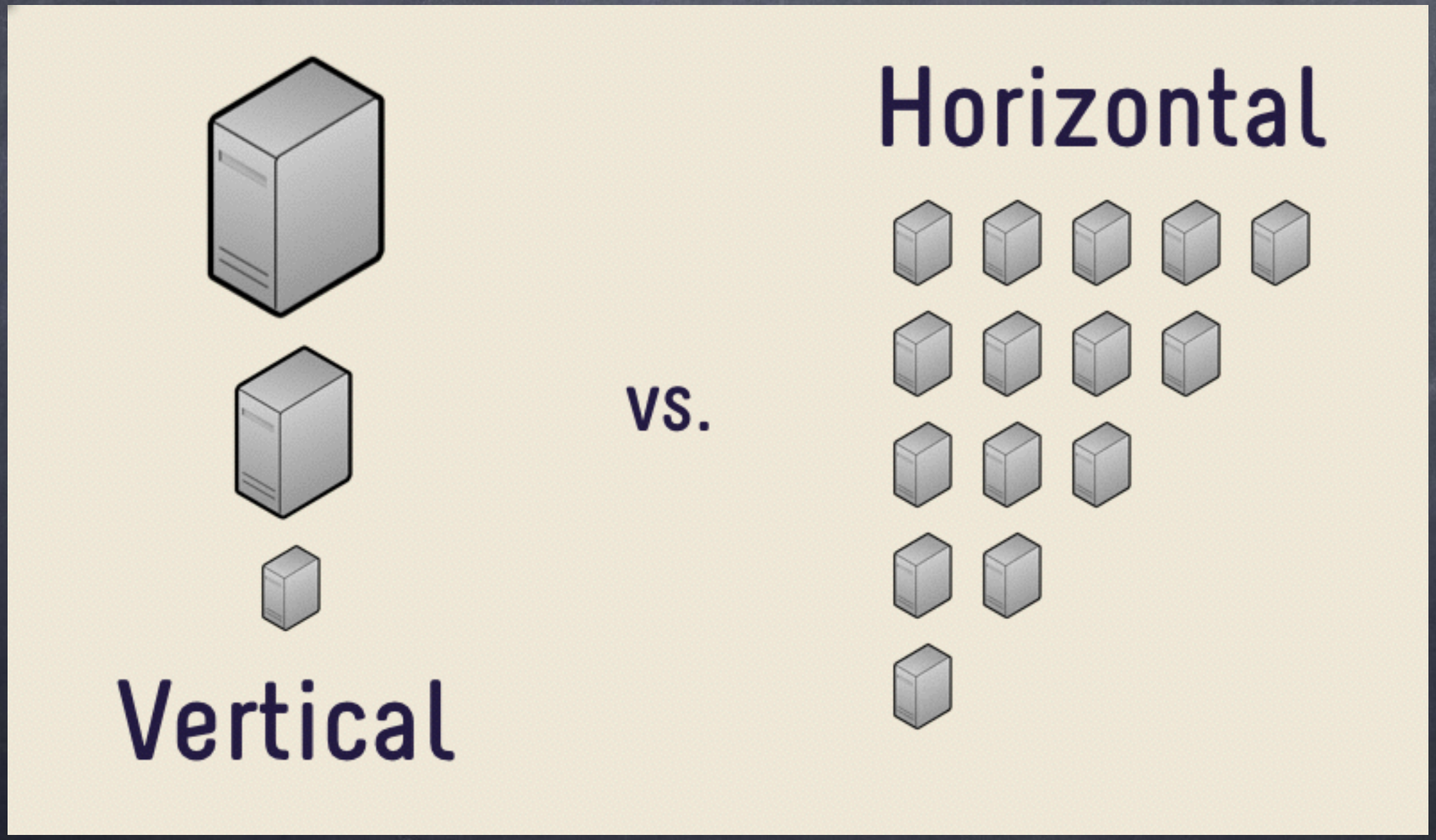
bigdata glossary

(python 으로 5분 만에 만듬)

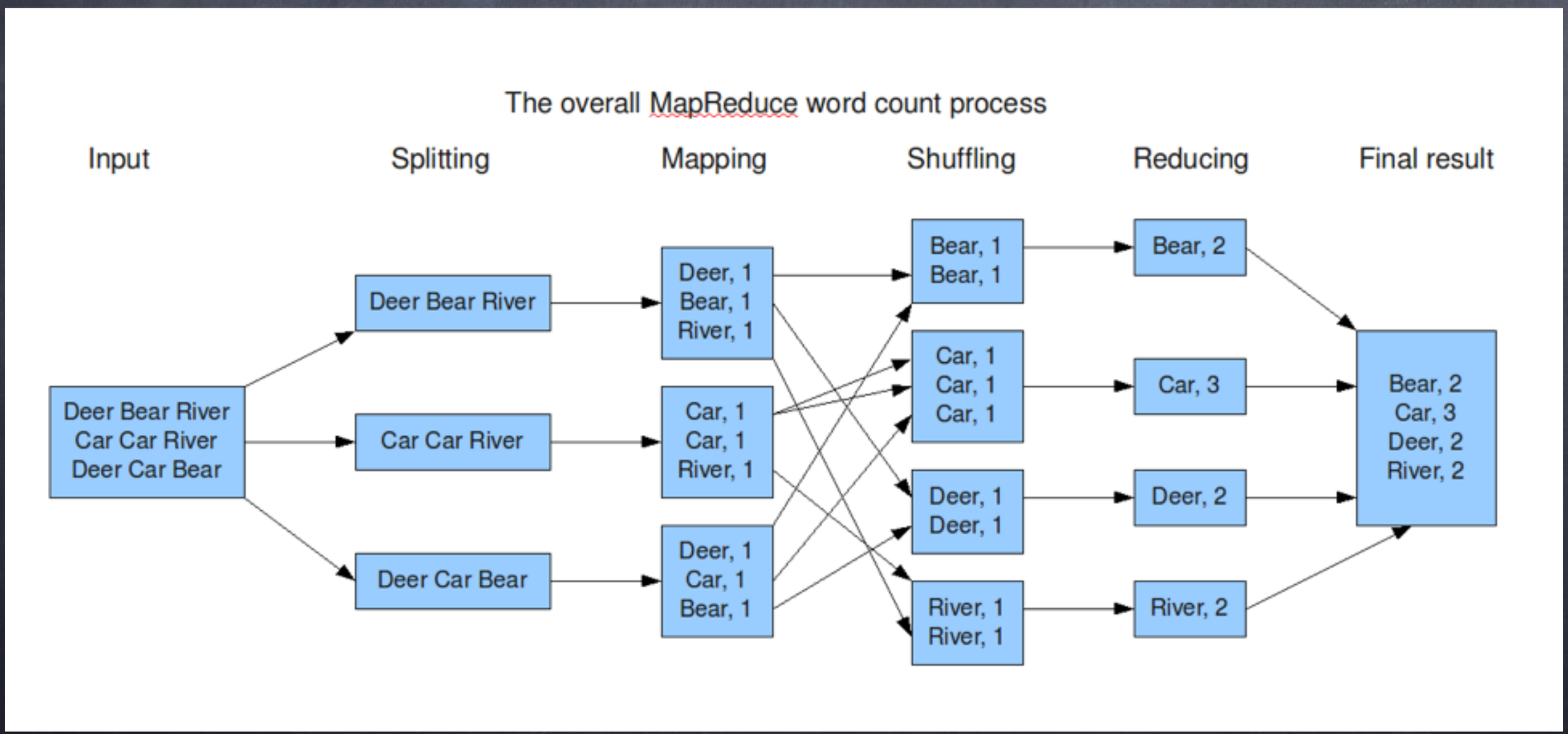


what is map reduce

수직 vs 수평



wordcount is hello world of Map Reduce

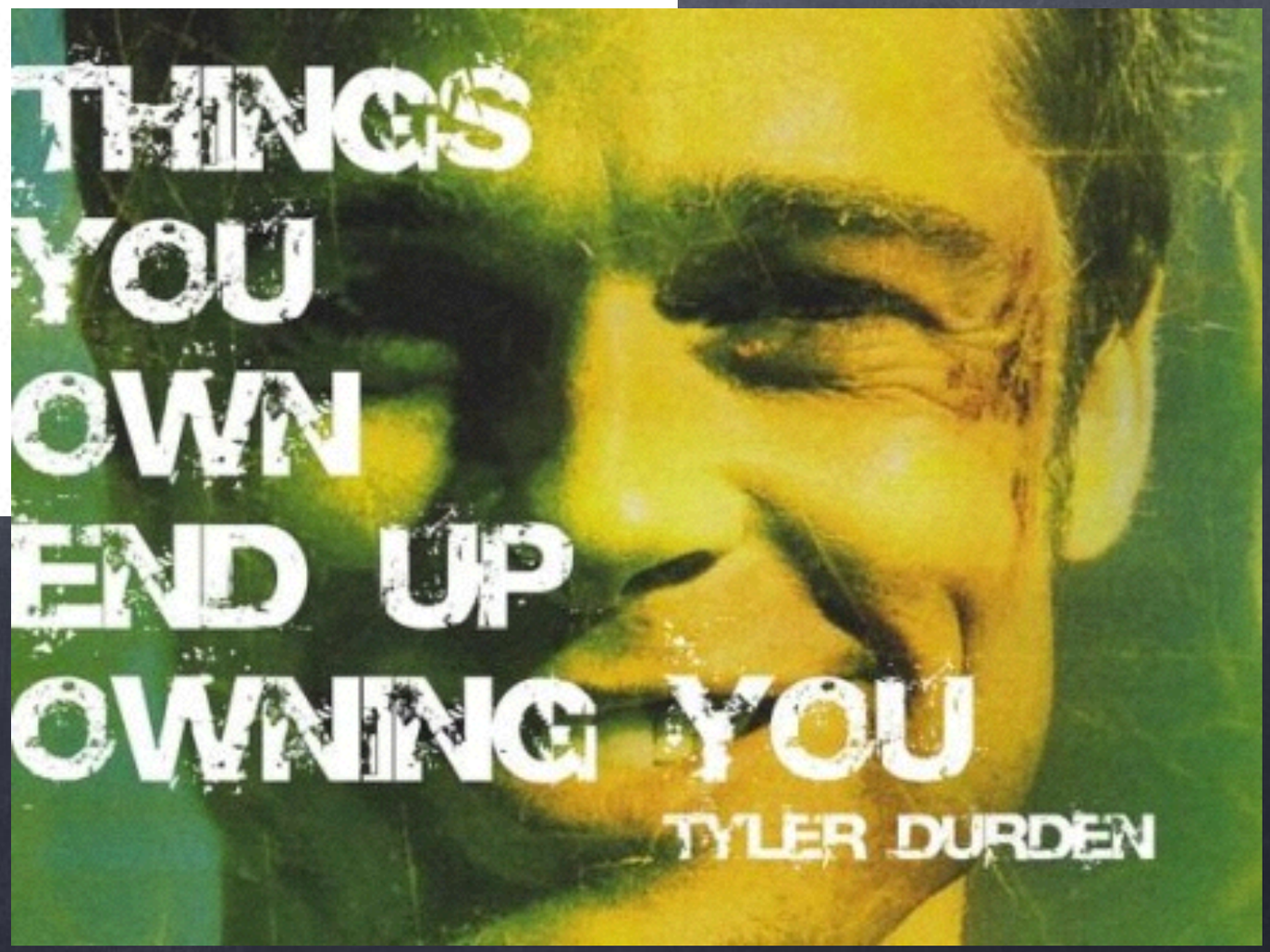


MapReduce 구현

@ hadoop
 - 하둡 스트리밍을 이용, 여러 언어에서 사용할 수 있음
- 사실상의 표준
- 수 많은 오픈소스 프로젝트, 보조도구들
- 하지만 자바

왜 하둡/자바 디스하나요?

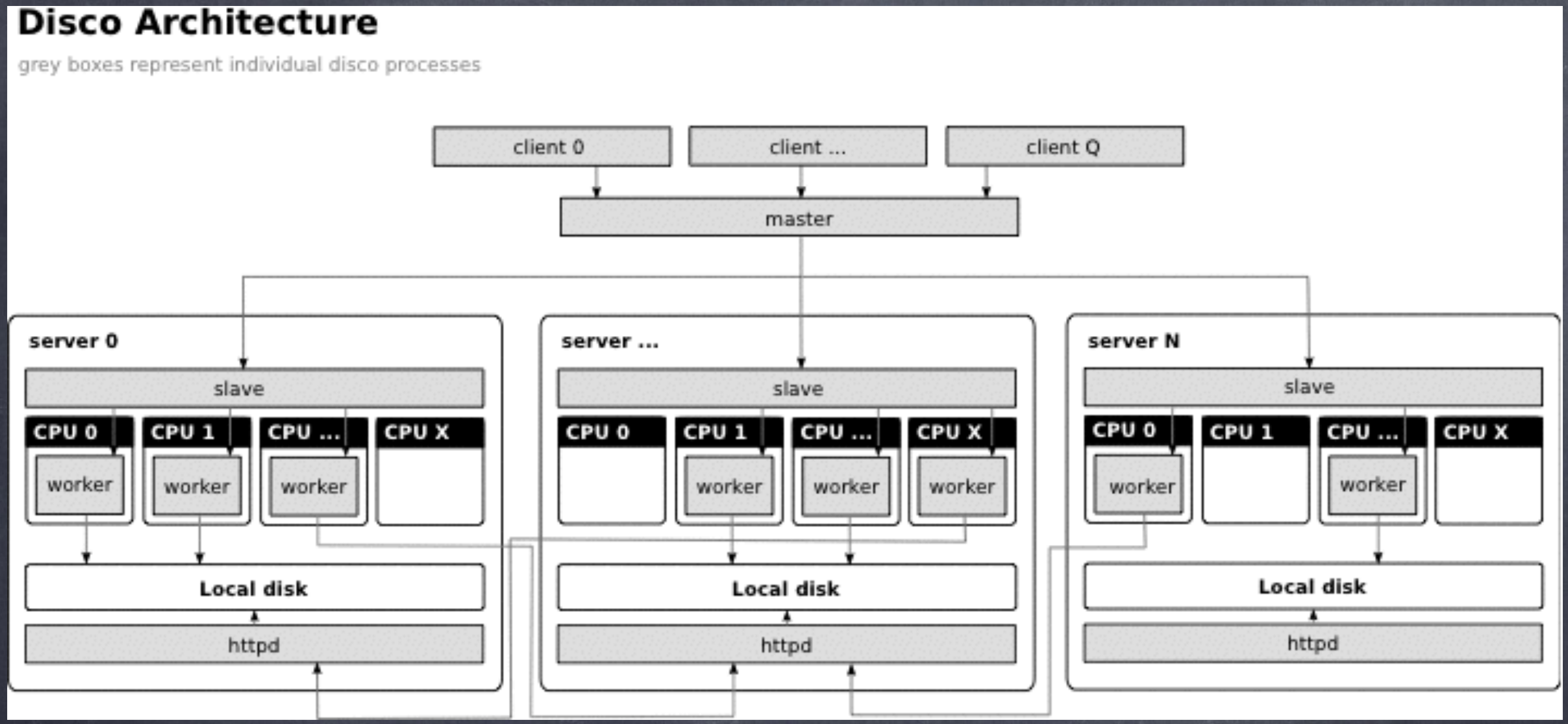

WE BUY THINGS WE DON'T NEED
WITH MONEY WE DON'T HAVE
TO IMPRESS PEOPLE WE DON'T LIKE.


THINGS
YOU
OWN
END UP
OWNING YOU
TYLER DURDEN

introducing disco

disco

@ erlang + python
@ 웹UI와 잡 관리는 erlang
@ 나머지는 전부 python
@ 하둡보다 매우 짧은 소스코드(10배?)
@ 설치가... 매우 쉬움
@ worker protocol
@ 노키아에서 ville tuulos 가 시작

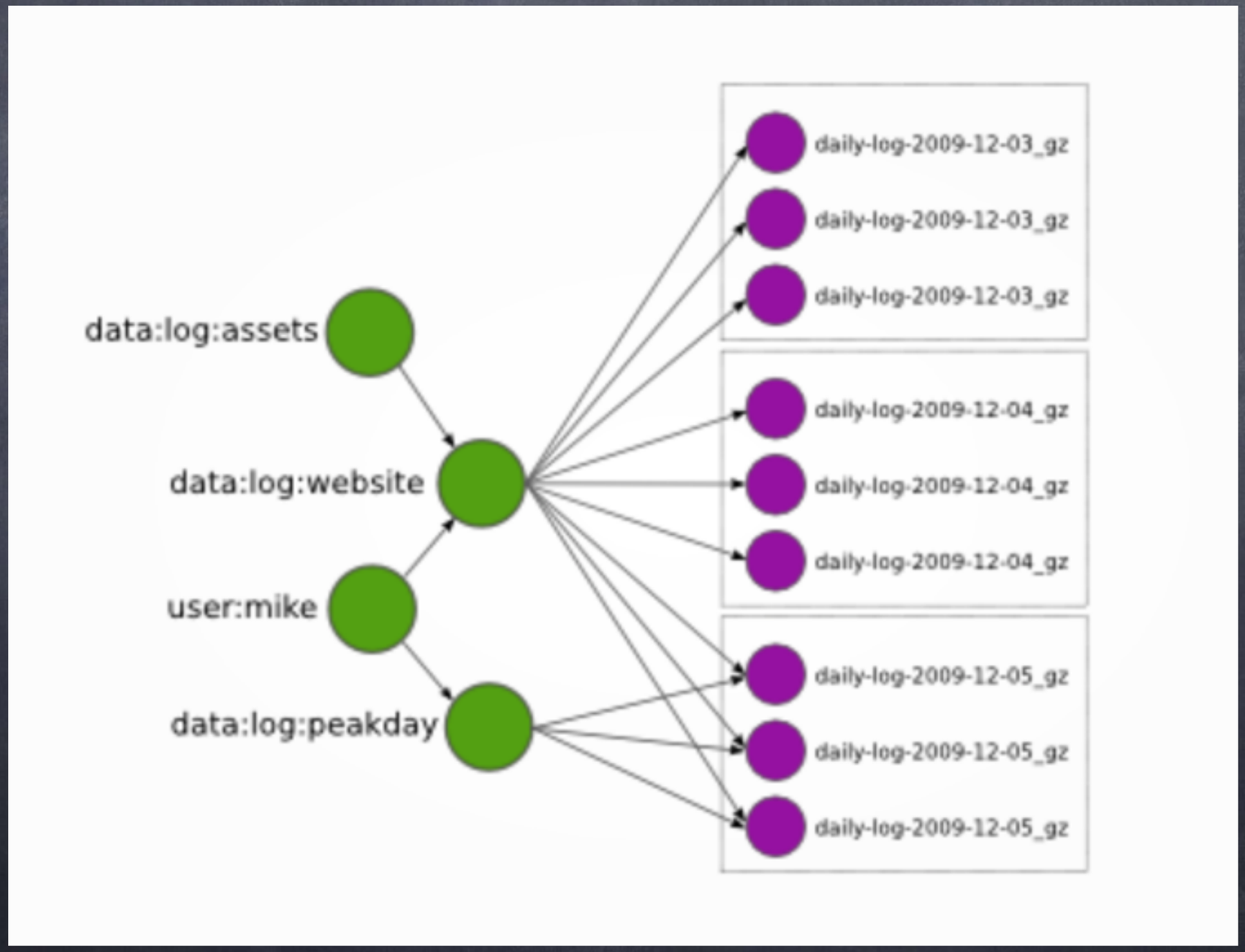


client master slave worker

- Disco users start Disco jobs in Python scripts.
- Jobs requests are sent over HTTP to the master.
- Master is an *Erlang* process that receives requests over HTTP.
- Master launches *slaves* on each node over *SSH*.
- Slaves run Disco tasks in *worker* processes.

DDFS

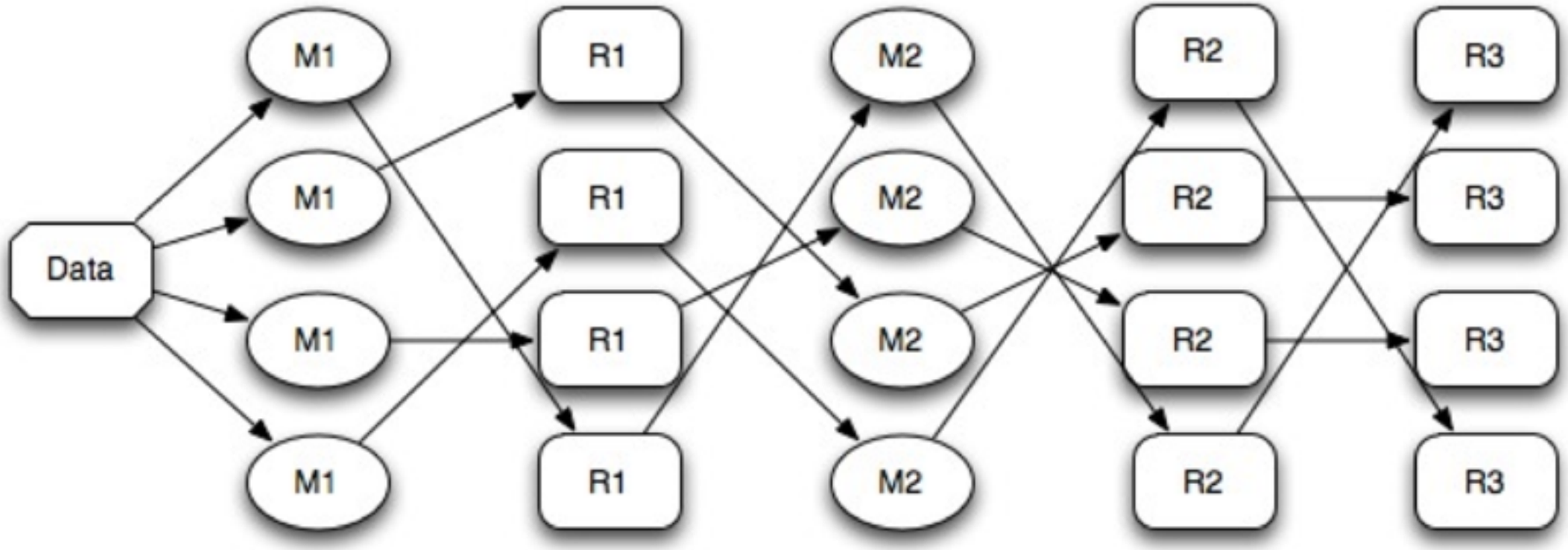
@ tag based file system



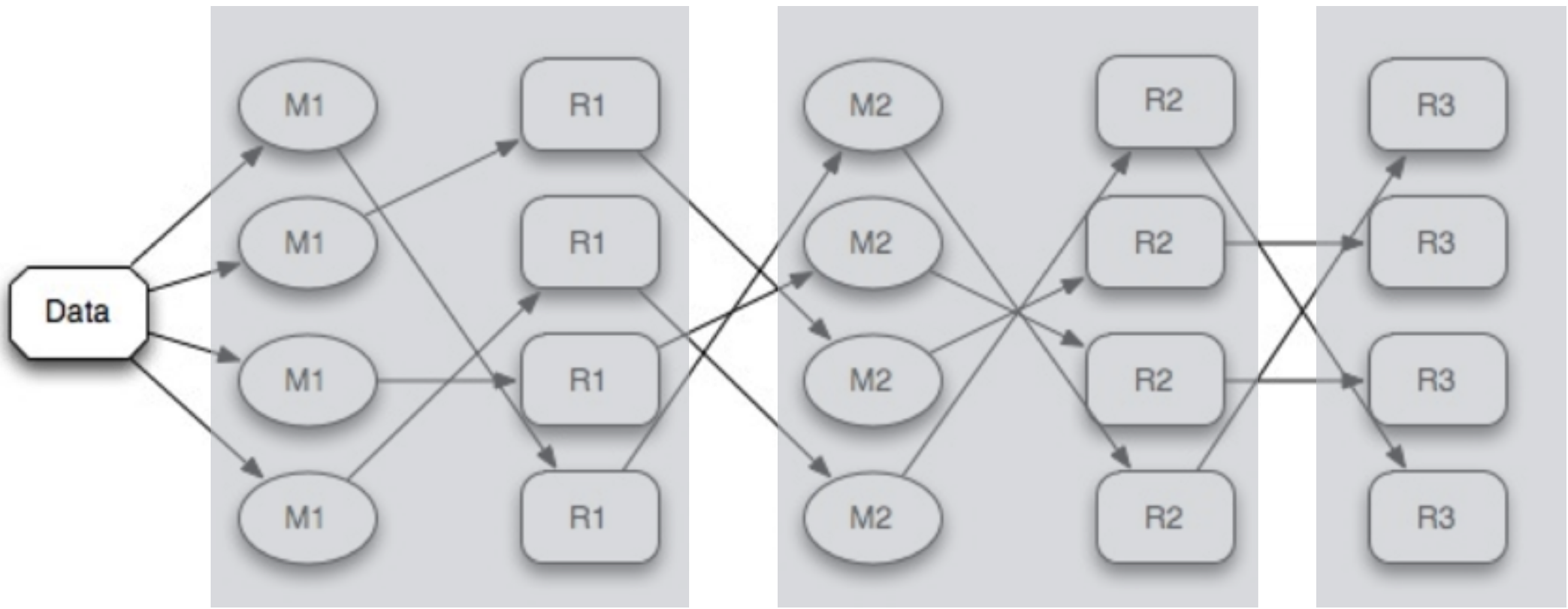
DDFS

```
$ ddfs chunk python:grail:script ./holy-grail.txt
created: disco://compute-0-2/ddfs/vol0/blob/1f/holy-grail_txt-0$516-af4c8-1a3db
disco://compute-0-0/ddfs/vol0/blob/1f/holy-grail_txt-0$516-af4c8-1a3db disco://compute-
0-3/ddfs/vol0/blob/1f/holy-grail_txt-0$516-af4c8-1a3db

$ ddfs blobs python:grail:script
disco://compute-0-2/ddfs/vol0/blob/1f/holy-grail_txt-0$516-af4c8-1a3db
disco://compute-0-0/ddfs/vol0/blob/1f/holy-grail_txt-0$516-af4c8-1a3db
disco://compute-0-3/ddfs/vol0/blob/1f/holy-grail_txt-0$516-af4c8-1a3db

$ ddfs xcat python:grail:script
Sacred-Texts  Legends and Sagas
Note: this is a transcript of the movie produced by an anonymous fan. Obviously the
original is copyrighted and anyone attempting to exploit this file commercially without
permission of Monty Python is a looney...--sacred-texts editor
```

chain jobs



chain jobs



헬로 월드

```
from disco.core import Job, result_iterator

def map(line, params):
    for word in line.split():
        yield word, 1

def reduce(iter, params):
    from disco.util import kvgroup
    for word, counts in kvgroup(sorted(iter)):
        yield word, sum(counts)

if __name__ == '__main__':
    job = Job().run(input=["http://discoproject.org/media/text/chekhov.txt"],
                    map=map,
                    reduce=reduce)
    for word, count in result_iterator(job.wait(show=True)):
        print(word, count)
```

build your own cluster

내 노트북이 리눅스다.

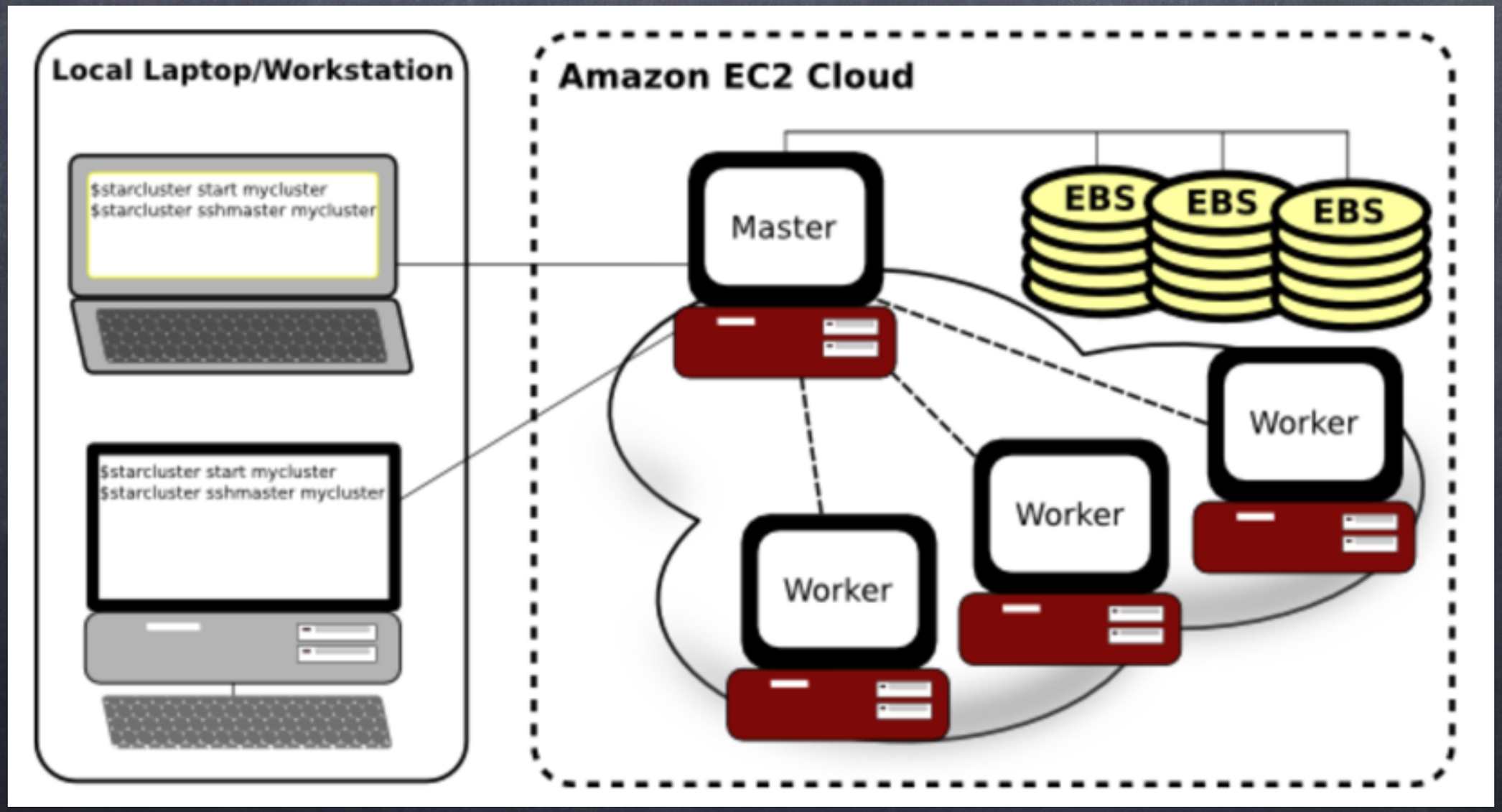
@ 공식 싸이트에서 Setting up Disco 따라하면 5분

내 노트북이 맥북이다.

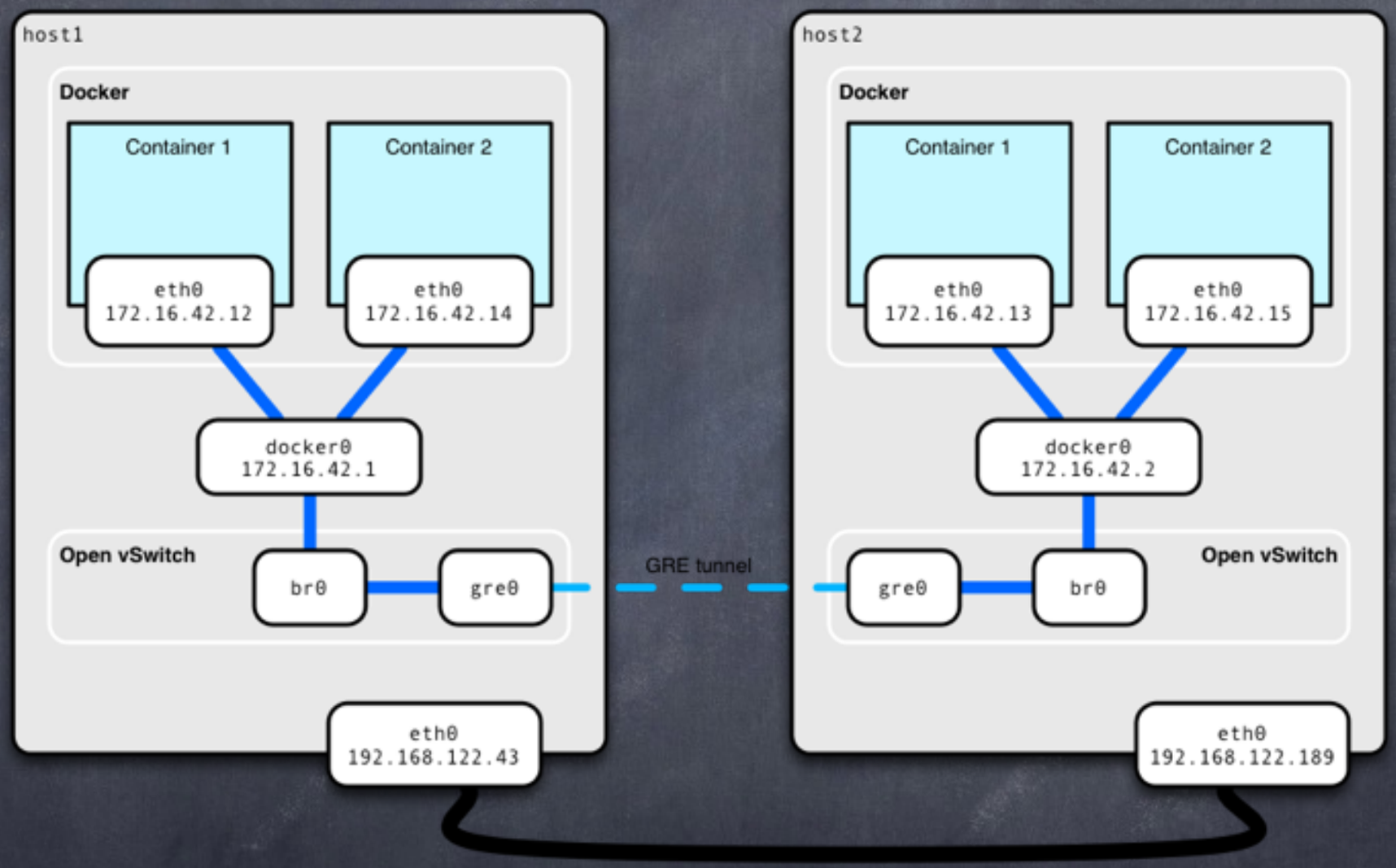
@ 공식 싸이트에서 Setting up Disco 따라하면 5분
@ erlang crash dump 생기면 1주일
@ 그냥 vagrant 까세요
@ slave 노드에 pycurl 필요함
@ hostname 안 맞추면 아무것도 안됨

Amazon EC2

@ starcluster : python !!



docker 시도: 실패



(성공하신 분 개인적으로 연락주시면 후사하겠습니다)

홈 클러스터

disco status

status | configure

imac

64.2GB

53	0	0

macmini

59.1GB

44	0	0

macpro

45.9GB

552	0	24

mbp15

84.3GB

0	56	0

SortJob@581:4bc

DiscoAPIJob@581:4

ConfigJob@581:4

ConfigJob@581:4bd

Job@581:4bd17:

```
[ ID] Interval           Transfer     Bandwidth
[  7]   0.00-10.00  sec  9.98 GBytes  8.57 Gbits/sec
[  7]   0.00-10.00  sec  9.98 GBytes  8.57 Gbits/sec

iperf Done.
┌─[kdog@KDOGMBPR15] - [~] - [2014-08-09 12:35:44]
└─[0] <> iperf3 -c 192.168.2.15
Connecting to host 192.168.2.15, port 5201
[  7] local 192.168.2.10 port 50735 connected to 192.168.
[ ID] Interval           Transfer     Bandwidth
[  7]   0.00-1.00   sec   806 MBytes  6.76 Gbits/sec
[  7]   1.00-2.00   sec   810 MBytes  6.79 Gbits/sec
[  7]   2.00-3.00   sec   799 MBytes  6.70 Gbits/sec
```

생각

@ erlang 은 또 하나의 진입장벽이다
@ 왠만한 데이터는 DBMS가 답임
@ HBase Pig Hive 등을 만들어서 쓴다면?
 @ MR을 더욱 효과적으로?
  -> 리액 node에 slave-worker 를 띄울 수 있다면
 @ MR은 전체 빅데이터 의사결정의 일부분으로 설계해야
 @ 맥 프로 정말 좋음

Where to Start?

@ discoproject.org(massive data, minimal code)

@ run tests(disco/tests)

-> run_tests_python25

@ follow for

-> scipy conference

-> ville tuulos(founder of disco)

@ NoSQL distilled(MR 설명, 인사이트에서 번역 나와있음ㅎ)

감사합니다!!