

*here be
grouperfish*



Grouperfish
clustering engine

Grouperfish clustering engine

client

client

...

↓ *PUT docs*

↓ *GET clusters*

REST node
nodeJS

REST node
nodeJS

...

service layer
req handling

PUT docs

GET clusters

OFFER doc

storage node



riak



redis

storage node



riak



redis

...

data layer
storage & indexing



task queue

RabbitMQ

POLL doc

PUT cluster



worker node
mahout & jetty



worker node
mahout & jetty

...

processing layer
scheduling
clustering (*small collections*)

↓ *APPEND docs*

↓ *GET clusters*

hadoop

hadoop

...

batch layer
clustering (*large collections*)

Riak contents

```
{
```

```
"some-ns/docs/some-doc-id": {  
  text: "I am a document 2 be clustrd.",  
  clusters: {  
    "collection-key-x": "label-in-x",  
    "collection-key-y": "label-in-y",  
    ...  
  }  
},
```

for updates & reconstruction of Redis

```
"some-ns/dictionary/some-collection-key":  
  /* binary dictionary file for Mahout */,  
  
"some-ns/vectors/some-collection-key":  
  /* binary vectors for Mahout */  
  
"some-ns/centroids/some-collection-key":  
  /* binary previous cluster centroids for Mahout */
```

for Mahout. Might want to just use HDFS instead

```
...  
}
```

Redis contents

```
{
```

GET requests

```
"clusters/some-ns/some-collection-key":  
  ["label-1", "label-2", ..., "label-k"],  
  
"cluster/some-ns/some-collection-key/label-1":  
  ["doc-id-1", "doc-id-2", ..., "doc-id-n"],
```

```
"size/some-ns/some-collection-key":  
  7185,  
  
"lock/some-ns/some-collection-key":  
  "pwn3d",  
  
"new/some-ns/some-collection-key":  
  ["doc-id-1", "doc-id-2", ...]
```

```
}
```

scheduling & updates