here be
grouperfish

Grouperfish

clustering engine

Grouperfish clustering engine

client client ...

PUT docs

GET clusters

REST node
nodeJS

REST node
nodeJS

...

PUT docs GET clusters

OFFER doc

storage node
redis
HBASE

storage node
redis
HBASE

...

task queue
RabbitMQ

POLL doc

PUT cluster

worker node
mahout & jetty

worker node
mahout & jetty

...

APPEND docs

GET clusters

hadoop hadoop ...

service layer
request handling

data layer
storage & indexing

processing layer
scheduling
clustering *(small collections)*

batch layer
clustering *(large collections)*

Hbase contents

top-keys = record keys, inner keys = column qualifiers

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

for updates & reconstruction of Redis

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

  "some-ns/some-collection-key/vectors":
    /* binary vectors for Mahout */

  "some-ns/some-collection-key/centroids":
    /* binary previous cluster centroids for k-means */
```

for Mahout

```
  …
}
```

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",

  “last-updated/some-ns/some-coll…”:
     “2011-03-08T13:56.000Z”

  "new/some-ns/some-collection-key":
     ["doc-id-1", "doc-id-2", …]
```

scheduling & updates

```
}
```