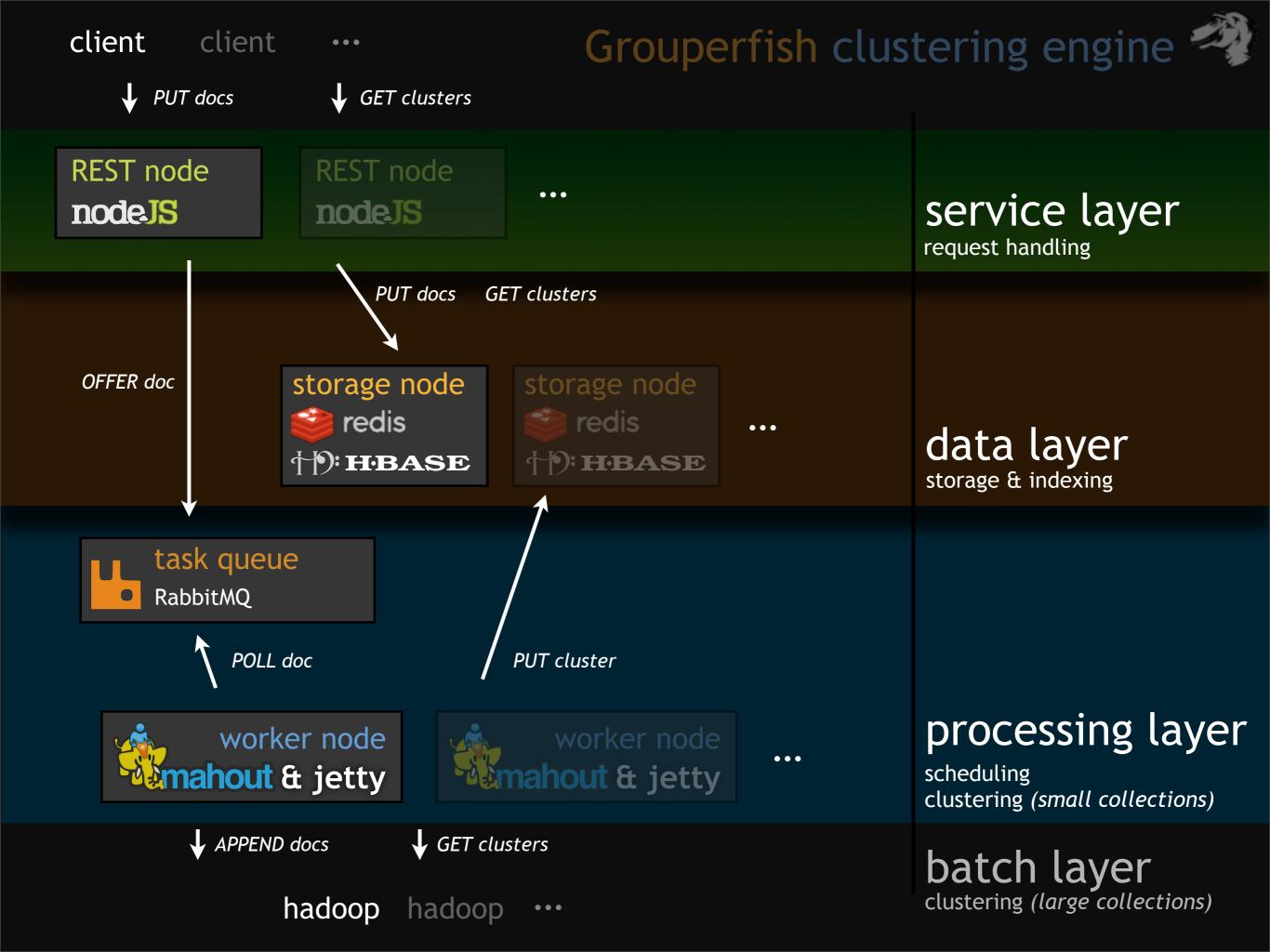
here be grouperfish





Grouperfish clustering engine



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

```
"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"],
```

scheduling & updates