ElasticUtils Documentation

Release dev

Mozilla Foundation

October 20, 2015

Contents

1 Project **3**
- 1.1 What's new in ElasticUtils . . . 3
- 1.2 Elasticsearch theory . . . 14
- 1.3 Resources . . . 15

2 User's Guide **17**
- 2.1 Installation . . . 17
- 2.2 Indexing . . . 17
- 2.3 Mapping types and Indexables . . . 22
- 2.4 Searching . . . 24
- 2.5 More like this: `MLT` . . . 40
- 2.6 Debugging . . . 40
- 2.7 API docs . . . 42
- 2.8 Migrating from Elasticsearch 0.90 to 1.x with ElasticUtils . . . 43

3 Using ElasticUtils with Django **47**
- 3.1 Using ElasticUtils with Django . . . 47
- 3.2 Django API docs . . . 51

4 Contributor's Guide **53**
- 4.1 Join this project! . . . 53
- 4.2 Hacking HOWTO . . . 53
- 4.3 Conventions . . . 54
- 4.4 Documentation . . . 54
- 4.5 Running and writing tests . . . 54
- 4.6 Release process . . . 55

5 Sample programs **57**
- 5.1 Basic sample program . . . 57
- 5.2 Sample program using facets . . . 59

6 Indices and tables **63**

Python Module Index **65**

Deprecated January 5th, 2015

This project is no longer being maintained. Last release is ElasticUtils 0.10.2. You should consider switching to elasticsearch-dsl-py.

Version dev

Code https://github.com/mozilla/elasticutils

License BSD; see LICENSE file

Issues https://github.com/mozilla/elasticutils/issues

Documentation http://elasticutils.readthedocs.org/

IRC #elasticutils on irc.mozilla.org

ElasticUtils is a Python library that gives you a chainable search API for Elasticsearch as well as some other tools to make it easier to integrate Elasticsearch into your application.

So what's it like? Let's do a couple basic things:

Create an instance of `elasticutils.S` and tell it which index and doctype to look at.

```
>>> from elasticutils import S, F
>>> s = S().indexes('blog-index').doctypes('blog-entry')
```

Print the count of everything in that index with that type:

```
>>> s.count()
4
```

Show titles of all blog entries with "elasticutils" in the title:

```
>>> s = s.query(title__match='elasticutils')
>>> [result['title'] for result in s]
[u'ElasticUtils v0.4 released!', u'elasticutils status -- May 18th, 2012',
u'ElasticUtils sprint at PyCon US 2013']
```

You can also use properties rather than keys:

```
>>> [result.title for result in s]
[u'ElasticUtils v0.4 released!', u'elasticutils status -- May 18th, 2012',
u'ElasticUtils sprint at PyCon US 2013']
```

Filter out entries related to PyCon:

```
>>> s = s.filter(~F(tag='pycon'))
>>> [result['title'] for result in s]
[u'ElasticUtils v0.4 released!', u'elasticutils status -- May 18th, 2012']
```

Show only the top result:

```
>>> s = s[:1]
>>> [result['title'] for result in s]
[u'ElasticUtils v0.4 released!']
```

That's the gist of it!

CHAPTER 1

Project

1.1 What's new in ElasticUtils

- *Version 0.11: There will be no 0.11*
- *Version 0.10.3: March 4th, 2015*
- *Version 0.10.2: October 10th, 2014*
- *Version 0.10.1: September 22nd, 2014*
- *Version 0.10: August 19th, 2014*
- *Version 0.9.1: April 7th, 2014*
- *Version 0.9: April 3rd, 2014*
- *Version 0.8.2: January 6th, 2014*
- *Version 0.8.1: September 13th, 2013*
- *Version 0.8: August 19th, 2013*
- *Version 0.7: Released June 12th, 2013*
- *Version 0.6: Released January 17th, 2013*
- *Version 0.5: Released September 4th, 2012*
- *Version 0.4: Released July 31st, 2012*
- *Version 0.3: Released June 1st, 2012*

1.1.1 Version 0.11: There will be no 0.11

Warning: Development on this project has ceased. There will be no 0.11.

1.1.2 Version 0.10.3: March 4th, 2015

Changes:

- Added support for terms_stats facet.

1.1.3 Version 0.10.2: October 10th, 2014

Note: This has been tested with Elasticsearch 0.90 up through 1.3.4. We don't support versions of earlier than 0.90. This supports elasticsearch-py >= 1.0.

This is a bridging release to help people migrate from Elasticsearch <= 0.90 to Elasticsearch >= 1.0.

The next version of ElasticUtils will not support versions of Elasticsearch < 1.0.

Changes:

- Fixed monkeypatch to work with all bulk op_types (e.g. insert, create, update and delete)

1.1.4 Version 0.10.1: September 22nd, 2014

Note: This supports Elasticsearch 0.90, 1.0, 1.1 and 1.2. It doesn't support versions earlier than 0.90 or later than 1.2.

This supports elasticsearch-py >= 1.0.

This is a bridging release to help people migrate from Elasticsearch <= 0.90 to Elasticsearch >= 1.0.

The next version of ElasticUtils will not support versions of Elasticsearch < 1.0.

API-breaking changes:

- **requires elasticsearch-py >= 1.0**

Changes:

- Add distance filter
- Fix tests to work with Elasticsearch 1.2
- Invert monkeypatch for bulk indexing
- Fix infinite recursion when pickling MappingType instances
- Invert monkeypatch–ElasticUtils now requires elasticsearch-py >= 1.0

1.1.5 Version 0.10: August 19th, 2014

Note: This version supports Elasticsearch 0.90, 1.0 and 1.1. It does not support versions earlier than 0.90 or later than 1.1.

ElasticUtils 0.10 does not work with elasticsearch-py > 0.4.5.

This is a bridging release to help people migrate from Elasticsearch <= 0.90 to Elasticsearch >= 1.0.

The next version of ElasticUtils will not support versions of Elasticsearch < 1.0.

API-breaking changes:

- **big ‘‘.values_list()‘‘ and ‘‘.values_dict()‘‘ changes**

  `.values_list()` and `.values_dict()` will now **always** specify the Elasticsearch `fields` property.

  If you call these two functions with no arguments (i.e. you specify no fields), they will send `fields=*` to Elasticsearch. It will send any fields marked as stored in the document mapping. If you have no fields marked as stored, then it will return the id and type of the result.

  If you call these two functions with arguments (i.e. you specify fields), then it'll return those fields—same as before.

  However, they now return all values as lists. For example:

```
>>> S().values_list()
[([100], ['bob'], [40]), ...]

>>> S().values_list('id')
[([100],), ([101],), ...]

>>> S().values_dict()
[({'id': [100], 'name': ['bob'], 'weight': [40]}), ...]

>>> S().values_dict('id', 'name')
[({'id': [100], 'name': ['bob']}), ...]
```

- Removed `text` and `text_phrase` queries. They're renamed in Elasticsearch to `match` and `match_phrase`.
- Removed `startswith` query. Replace uses of it with `prefix`.

Changes:

- Python 3 support (Python >= 3.3)
- Supports Elasticsearch 0.90, 1.0 and 1.1

1.1.6 Version 0.9.1: April 7th, 2014

Changes:

- **Fixed bug with facets that are both sized and filtered**

1.1.7 Version 0.9: April 3rd, 2014

Note: This is a *big* change. We switched from pyelasticsearch to elasticsearch-py. The Elasticsearch object you get back from `get_es` is pretty different. When you upgrade to ElasticUtils v0.9, you'll probably need to rewrite code.

If this terrifies you, read through these notes carefully and/or stay with ElasticUtils v0.8.

Note: This version supports Elasticsearch 0.20 through 0.90. It does not yet support Elasticsearch 1.0. Support for 1.0 and later will be in a later version of ElasticUtils.

API-breaking changes:

- **elasticsearch-py >= v0.4.3 and < 1.0 required.**

  ElasticUtils now uses elasticsearch-py.

  Note: You have to use elasticsearch-py >= v0.4.3 and < 1.0. ElasticUtils does not support elasticsearch-py 1.0. Support for later versions will come in a future ElasticUtils release.

- **pyelasticsearch no longer needed**

  You can remove pyelasticsearch and its requirements.

- **thrift now supported!**

  elasticsearch-py supports http, thrift and memcache protocols, so you can use any of them now.

- **pyelasticsearch -> elasticsearch-py changes.**

  If you called `elasticutils.get_es()` and got a pyelasticsearch *ElasticSearch* object and did things with that (create index, create mappings, delete indexes, indexing, cluster health, ...), you're going to need to make some changes.

  You can either:

  1. rewrite that code to use elasticsearch-py *Elasticsearch* equivalents, or
  2. write a different function that returns a pyelasticsearch *ElasticSearch* object and use that

  Rewriting shouldn't be too hard. The elasticseary-py documentation is pretty good and for most things, there's a 1-to-1 translation. Also, in many cases it's cleaner, so you'll probably be removing code.

- **S.all() no longer returns all results**

  If you were using *S.all()* to return all search results, you should change it to *S.everything()*.

- **'S._build_query()' was changed to 'S.build_search()'**

  This makes the method public and also changes the name and documentation to be more correct.

  If you really need a *S._build_query()*, add it to your *S* subclass.

- **Search results metadata is now in the 'es_meta' object**

  Previously, you would access search results metadata like this:

```
obj._id
obj._highlight
obj._score

etc.
```

  In order to make those accessible in Django templates, we moved them into an es_meta object. You can now access them like this:

```
obj.es_meta.id
obj.es_meta.highlight
obj.es_meta.score

etc.
```

Changes:

- **added S.everything() which does what S.all() did**
- **index_objects celery task can now take es and index args**
- **unindex_objects celery task can now take es and index args**
- **added S.suggestions() support**
- **added S.query_and_fetch() support**
- **added S.search_type()**
- **S.facet() can now take a size keyword argument**
- **S.facet_couts() now returns a dict of FacetResults objects**

  The FacetResults object contains all the data we get back from that section in the Elasticsearch response.

- **SearchResults now has facet data in facets property**

- **elasticutils.estestcase.ESTestCase available and cleaned up**

  Previously, it was in `elasticutils/tests/__init__.py`. This makes it so everyone can use the same TestCase subclass we're using for our tests.

1.1.8 Version 0.8.2: January 6th, 2014

Changes:

- **Allow pyelasticsearch 0.6.1.**

  This alleviates part of the problem in issue #163.

- **Add tox.ini file.**

  We're testing with Python 2.6 and 2.7 on Django 1.4, 1.5 and 1.6.

- **Add caching for empty results.**

  ElasticUtils will now correctly remember when it got no results from a search and won't redo the search.

- **Add support for query and filter facets.**

- **Attach facets to search result objects.**

- **order_by() accepts a dict as the sort field so you can do advanced sorts.**

1.1.9 Version 0.8.1: September 13th, 2013

API-breaking changes:

- **Indexable.index overwrite_existing argument default changed**

  In v0.8, we added the `overwrite_existing` argument, but made it default to False. That's different than what pyelasticsearch does.

  In v0.8.1, we changed the default to True which is in line with what pyelasticsearch does.

  If you were depending on the old behavior, then you need to update your indexing code to set `overwrite_existing=False`.

1.1.10 Version 0.8: August 19th, 2013

API-breaking changes:

- **pyelasticsearch v0.6 or later now required.**

  Further, since pyelasticsearch has released versions that aren't backwards compatible, we're now pegging on specific versions.

- **celery 2.5.5 or later now required.**

  You can ignore this if you're not using the Django celery tasks.

- **Indexable.index arguments changed**

  pyelasticsearch changed arguments, so we did, too. We dropped the `force_insert` argument (which wasn't working) and picked up `overwrite_existing`.

  `overwrite_existing` defaults to False which means it will **not** overwrite existing documents in the index.

  Note: This was a mistake since pyelasticsearch defaults to True. We changed this in 0.8.1.

Changes:

- **Added support for ‘‘range‘‘ queries and filters.**

  `range` is a nice shorthand for `gte` and `lte`.

- **S.filter_raw added**

  If `elasticutils.S.filter()` isn't doing as it's told, then you can skip it and use the Elasticsearch API to create the filter clause of the search by hand with `elasticutils.S.filter_raw()`.

- Moved requirements files to `requirements/`.

1.1.11 Version 0.7: Released June 12th, 2013

Note: This is a *big* change. We switched from pyes to pyelasticsearch. In doing that, we changed a handful of signatures, nixed some functionality that didn't make any sense any more, and cleaned a bunch of things up.

If this terrifies you, read through these notes carefully and/or stay with v0.6.

API-breaking changes:

- **pyelasticsearch v0.4 or later now required.**

  ElasticUtils now requires pyelasticsearch v0.4 or later and its requirements.

- **elasticutils.PYES_VERSION is removed.**

  Since we're not using pyes, we removed *elasticutils.PYES_VERSION*.

- **ElasticUtils no longer supports thrift.**

  Pretty sure we did a lousy job of supporting it before—it was all in the pyes code and we had no tests for it.

- **get_es() signatures have changed.**

  - takes urls now instead of hosts
  - dump_curl argument is now gone
  - default_indexes argument is gone

  The arguments correspond with pyelasticsearch *ElasticSearch* object.

  ElasticUtils uses HTTP urls for connecting to Elasticsearch now. Previously, you'd do:

```
get_es(hosts=['localhost:9200'])        # Old way
```

  Now you do:

```
get_es(urls=['http://localhost:9200'])  # New way
```

  The dump_curl argument was helpful for debugging, but we don't really need it anymore. See the *Debugging* for better debugging methods.

  Will now raise a *DeprecationWarning* if you pass in *hosts* argument.

- **S searches all indexes and doctypes by default.**

  Previously, if you did:

```
S()
```

it'd search an index named "default" for doctypes "document". That was dumb. Now it searches all indexes and all doctypes by default.

- **S.es_builder is gone.**

  `es_builder()` was there to get around problems with pyes' ES class. The pyelasticsearch *ElasticSearch* class is more straightforward, so we don't need to do circus shenanigans.

  You can probably do what you need to with either the `es()` transform or by subclassing *S* and overriding the `get_es()` method.

- **MLT arguments changed.**

  The *fields* argument in the constructor was renamed to *mlt_fields* to be in line with Elasticsearch API names.

  Will now raise a *DeprecationWarning* if you pass in *fields* argument.

- **MappingType get_indexes renamed to get_index.**

  MappingType had a method called *get_indexes*. This is now *get_index* because it should return a single index name.

- **Added Indexable mixin for indexing bits for MappingTypes.**

- **Django: changed settings.**

  Changed ES_HOSTS setting to ES_URLS. This is both a name and a value change. ES_URLS takes a list of strings each is an http url. You'll neex to update your settings files from:

```
ES_HOSTS = ['localhost:9200']         # Old way
```

  to:

```
ES_URLS = ['http://localhost:9200']   # New way
```

  ES_DUMP_CURL is gone.

- **Django: removed the statsd code.**

- **Django: ESTestCase was improved, documented and bugs squashed.**

  It was improved, documented and bugs were squashed. It's now used by the test suite.

- **Django: Indexable.index() method no longer has bulk argument.**

  The *Indexable.index()* method no longer does bulk indexing. The way pyes did this was kind of squirrely and caused issues if you didn't have the order of operations correct.

  Now *Indexable.index()* only indexes a single document.

  But wait...

- **Django: Indexable now has bulk_index().**

  pyes would keep track of all the things you wanted to bulk index and then at some point push them all. Instead of doing it under the hood, we added a separate *bulk_index()* method and now you control how many items get indexed in bulk in one pass.

- **Django: Indexable.refresh_index no longer takes a timeout argument.**

  pyelasticsearch *ElasticSearch.refresh* doesn't take a timesleep argument, so we don't need that anymore.

- **Django: Indexable es argument defaults to Indexable.get_es() now.**

  Previously it defaulted to *elasticsearch.contrib.django.get_es()*. Now it defaults to *Indexable.get_es()* class method making it more flexible.

- **Django: renamed DjangoMappingType to MappingType.**

- **Django: moved MappingType and Indexable.**

  They were in `elasticutils.contrib.django.models` and are now in `elasticutils.contrib.django`. Yay for slightly shorter module paths!

- **Django: ditched the cron module and its helpers.**

  It's not clear they ever worked (issue #21) and there are no tests.

- **pyes -> pyelasticsearch changes.**

  If you called `.get_es()` and got a pyes *ES* object and did things with that (create index, create mappings, delete indexes, indexing, cluster health, ...), you're going to need to make some changes.

  You can either:

  1. rewrite that code to use pyelasticsearch *ElasticSearch* equivalents, or
  2. write and use your own `get_es()` function that returns a pyes *ES* object

  Rewriting shouldn't be too hard. The pyelasticsearch documentation is pretty good and for most things, there's a 1-to-1 translation.

Changes:

- **pyes is no longer a requirement.**

  We no longer use pyes so you can remove it from your requirements.

- **S.execute added**

  This allows you to explicitly execute a search and get back a *SearchResults* instance.

  See `elasticutils.S.execute()` for details.

- **S.all added**

  Allows you to get **all** the search results possible rather than just the first 10 search results which is the default.

  You should consider using slices instead which allows you to specify the maximum number of results to get back.

  This is dangerous, so it's been documented with lots of warnings.

  See `elasticutils.S.all()` for details.

- **Added support for ''match'' and ''match_phrase'' queries.**

  Elasticsearch 0.19.9 renamed text query to match query. This adds support for `match` and `match_phrase`.

  See *Queries: query* for details.

- **Added support for ''wildcard'' and ''terms'' queries.**

  See *Queries: query* for details.

- **Reimplemented filter and query implementation.**

  The new implementations allow you to add handling for filters and queries that ElasticUtils doesn't handle as well as override what ElasticUtils does.

  See `elasticutils.S` for details.

- **S.query_raw added**

  If `elasticutils.S.query()` is getting you down, then you can skip it and use the Elasticsearch API to create the query clause of the search by hand with `elasticutils.S.query_raw()`.

- **Django: es_required_or_50x handles different exceptions.**

  Previously it handled:

  - pyes.urllib3.MaxRetryError
  - pyes.exceptions.IndexMissingException
  - pyes.exceptions.ElasticSearchException

  We're not using pyes anymore, so now it handles:

  - pyelasticsearch.exceptions.ConnectionError
  - pyelasticsearch.exceptions.ElasticHttpError
  - pyelasticsearch.exceptions.ElasticHttpNotFoundError
  - pyelasticsearch.exceptions.InvalidJsonResponseError
  - pyelasticsearch.exceptions.Timeout

  You probably don't need to do anything about this, but it's good to know.

- **Django: celery tasks rewritten.**

  The celery tasks were rewritten, docs were updated, and tests were added so they work now.

1.1.12 Version 0.6: Released January 17th, 2013

API-breaking changes:

- **S.values_dict no longer always includes id.**

  `values_dict` no longer always includes an 'id' field in the fields list if you don't specify it.

  Specifying no fields now returns **all** fields:

```
S().values_dict()
```

  Specifying fields now returns only those fields:

```
S().values_dict('name', 'number')
```

- **S.values_list no longer always includes id.**

  `values_list` no longer always includes an 'id' field in the fields list if you don't specify it.

  Specifying no fields now returns **all** fields:

```
S().values_list()
```

  Specifying fields now returns data for those fields in the order the fields are specified:

```
S().values_list('name', 'number')
```

- **Types have changed.**

  This is a big change.

  Up through ElasticUtils v0.5, *S* could take a type and that type was a model. This is now completely different.

  In ElasticUtils v0.6 and later, *S* takes a *MappingType*. A *MappingType* can be related to a model, but it itself should not be a model. This allows us to return search results as a list of *MappingType* instances which can do things rather than forcing you to do a db hit to get back instances that can do things.

This is similar to how django-haystack works with the SearchIndex class, except ElasticUtils doesn't yet support declarative mapping definition.

See documentation for more details.

- **By default, results are now DefaultMappingType.**

  In ElasticUtils v0.4 and v0.5, if the *S* was untyped and you didn't specify either `values_dict` or `values_list`, then the results would come back as a list of dicts.

  In ElasticUtils v0.5, if the *S* is untyped and you didn't specify either `values_dict` or `values_list`, then the results would come back as a list of *DefaultMappingType*.

  See documentation for more details.

- **elasticutils.contrib.django.models.SearchMixin is no more.**

  The *SearchMixin* class is replaced by *DjangoMappingType* which relates Elasticsearch mapping types to Django ORM models and *Indexable* which is a mixin that adds a bunch of index-related infrastructure.

Changes:

- Added `_source` and `_id` to the metadata decorated on the search results.

  See documentation for more details.

- Fixed `elasticutils.contrib.django.es_required_or_50x`.

  It works better now.

- prefix filter support.

  ElasticUtils supports prefix filters. You can do this now:

```
S().filter(name__prefix='odin')
```

1.1.13 Version 0.5: Released September 4th, 2012

API-breaking changes:

None.

Changes:

- Added `demote` transform: it adds boosting query support allowing you to do a negative query which reduces scores for documents that match.
- The elasticutils version is now available in `elasticutils.__version__` as well as `elasticutils._version.__version__`.
- Added `__in` support for queries. Doing:

```
S().query(foo__in=['a', 'b', 'c'])
```

  does a terms query now.

- Added *MLT* class which does morelikethis.
- Added API documentation for S, an index, `order_by` docs, fixed some icky bugs, and generally improved everything at least a little bit.

1.1.14 Version 0.4: Released July 31st, 2012

API-breaking changes:

- **ElasticUtils no longer requires Django.**

  If you're using Django, you should change your import statements from things like:

```
from elasticutils import get_es, S, F
```

  to:

```
from elasticutils.contrib.django import get_es, S, F
```

  Further, Django helper modules like `cron`, `tasks`, and `models` were all moved to `elasticutils.contrib.django`.

  We moved *ESTestCase* from `elasticutils.tests` to `elasticutils.contrib.django.estestcase`

  If you don't use Django, ElasticUtils is easier to use!

- **S no longer requires a type.**

  If you're not using Django, *S* no longer requires a type. If you don't specify a type, then ElasticUtils will return results as dicts.

- **Values and values_list changed.**

  `values()` was renamed to `values_list()`.

  `values_list()` (was `values()`) now always returns a list of tuples even if you only requested a single field. Previously, doing something like:

```
searcher = S().values_list('id')
```

  would return something like:

```
[1, 2, 3, 4, 5]
```

  Now it returns:

```
[(1,), (2,), (3,), (4,), (5,)]
```

- **Facet functionality was rewritten.**

  Changed `.facet()` to be arg-driven and allow for *filtered* and *global_* flags.

  Changed `.facets()` to `.facet_counts()` to match Django Haystack.

  Added `.facet_raw()` which allows you to do more complicated facets including scripting. This is similar to the original `.facet()` implementation.

Changes:

- Overhauled and cleaned up ElasticUtils tests. Running tests can be done with:

```
DJANGO_SETTINGS_MODULE=es_settings nosetests
```

- Default timeout was changed from 1 second to 5 seconds.
- Added `es` transform: it allows you to specify the settings with which to create an ES when the search is executed.
- Added `es_builder` transform: it allows you to specify a function that builds an ES which will be executed to create an ES when the search is executed.
- Added `indexes` transform: it allows you to specify the indexes to use for the search.

- Added `doctypes` transform: it allows you to specify the doctypes to use for the search.
- Added `explain` transform: it allows you to set the "explain" flag which gives you an explanation of how the score was calculated.

  I also added `elasticutils.utils.format_elasticutils` which formats the resulting explanation text into something slightly more readable. But it's likely this will change in the future.
- Added `boost` transform: it allows you to do query-time field boosting.
- Added support for `prefix`. It's the same as `startswith`, but it uses the same word that ElasticSearch uses. At some point, we'll remove support for `startswith`.
- Added support for `text_phrase` and `query_string` queries.
- Added `highlight` transform: generates highlighted fragments of content that matched the query.
- Removed requirement for nuggets.
- Continued to improve documentation.

1.1.15 Version 0.3: Released June 1st, 2012

Changes:

- Add documentation for debugging, project details and other things.
- Minor project cleanup to make it easier to maintain and use
- Make `get_es()` more useful. It now takes overrides that allow you to configure multiple kinds of ES objects for different purposes.

1.2 Elasticsearch theory

1.2.1 Indexes and types

Elasticsearch stores documents in an index allowing you to search them. The index is a container for documents. You can have multiple indexes in your cluster of Elasticsearch nodes.

Documents are typed. A type has a list of fields that are in the documents of that type. ElasticUtils calls this a "mapping type" or a "doc type" since the word "type" is somewhat ambiguous depending on the context.

See also:

http://www.elasticsearch.org/guide/reference/glossary/#index Elasticsearch explanation of indexes

http://www.elasticsearch.org/guide/reference/glossary/#mapping Elasticsearch explanation of mappings

http://www.elasticsearch.org/guide/reference/glossary/#type Elasticsearch explanation of types

1.2.2 Queries vs. filters

A search can contain queries and filters. The two things are very different.

A **filter** determines whether a document is in the results set or not. It doesn't affect scores. If you do a term filter on whether field *foo* has value *bar*, then the result set ONLY has documents where *foo* has value *bar*. Filters are fast and filter results are cached in Elasticsearch when appropriate. Use filters when you can.

A **query** affects the score for a document. If you do a term query on whether field *foo* has value *bar*, then the result set will score documents where the query holds true higher than documents where the query does not hold true. Queries are slower than filters and query results are not cached in Elasticsearch.

The other place where this affects things is when you specify facets. See *Facets* for details.

See also:

http://www.elasticsearch.org/guide/reference/query-dsl/ Elasticsearch Filters and Caching notes

1.3 Resources

1.3.1 Documentation

Elasticsearch guide

> **Elasticsearch documentation** http://www.elasticsearch.org/guide/
>
> **Elasticsearch 0.90 guide** http://www.elasticsearch.org/guide/en/elasticsearch/reference/0.90/index.html
>
> **Elasticsearch 1.x guide** http://www.elasticsearch.org/guide/en/elasticsearch/reference/current/index.html
>
> This is the canonical documentation.

elasticsearch-py documentation

> https://elasticsearch-py.readthedocs.org/en/latest/
>
> ElasticUtils sits on top of elasticsearch-py, so their documentation is very helpful.

1.3.2 Videos

Elasticsearch videos

> http://www.elasticsearch.org/videos/
>
> Lots of videos covering a variety of use cases and other things.

Elasticsearch video tutorials

> http://www.elasticsearch.org/tutorials/
>
> Covers deployment and using Elasticsearch for various things

FoodFightShow

> http://www.youtube.com/watch?v=dBWlXdmjjzY
>
> Covers Elasticsearch.

Erik Rose's talks:

> http://pyvideo.org/video/1784/elasticsearch-part-1-indexing-and-querying
>
> Elasticsearch provides an easy path to clusterable full-text search, with synonyms, faceting, and geographic math, but there's a paucity of written wisdom beyond its API docs. This talk, part 1 of a 2-part series, surveys its capabilities and shows how its internal data structures and algorithms work. With the groundwork laid, we explore how to choose efficient indexing and the right queries to make your apps go fast.

CHAPTER 2

User's Guide

2.1 Installation

2.1.1 Requirements

ElasticUtils requires:

- Python 2.6, 2.7, 3.3 or 3.4
- elasticsearch-py >= 1.0 and its dependencies
- Elasticsearch >= 0.90

  This does not work with versions of Elasticsearch older than 0.90.

2.1.2 Installation

There are a few ways to install ElasticUtils:

From PyPI

Do:

```
$ pip install elasticutils
```

From git

Do:

```
$ git clone git://github.com/mozilla/elasticutils.git
$ cd elasticutils
$ python setup.py install
```

2.2 Indexing

- *Overview*
- *Getting an Elasticsearch object*
- *Indexes*
- *Types and Mappings*
- *Indexing documents*
- *Deleting documents*
- *Refreshing*
- *Delete indexes*
- *Doing all of this with MappingTypes and Indexables*

2.2.1 Overview

ElasticUtils is primarily an API for searching. However, before you can search, you need to create an index and index your documents.

This chapter covers the indexing side of things. It does so lightly—for more details, read through the elasticsearch-py documentation and the Elasticsearch guide.

2.2.2 Getting an Elasticsearch object

ElasticUtils uses *elasticsearch-py* which comes with a handy *Elasticsearch* object. This lets you:

- create indexes
- create mappings
- apply settings
- check status
- etc.

To access this, you use `elasticutils.get_es()` which creates an *Elasticsearch* object for you.

See `elasticutils.get_es()` for more details.

See also:

http://elasticsearch-py.readthedocs.org/en/latest/api.html#elasticsearch elasticsearch-py Elasticsearch documentation.

2.2.3 Indexes

An *index* is a collection of documents.

Before you do anything, you need to have an index. You can create one with *.indices.create()*.

For example:

```
es = get_es()
es.indices.create(index='blog-index')
```

You can pass in settings, too. For example, you can set the refresh interval when creating the index:

```
es.indices.create(index='blog-index', body={'refresh_interval': '5s'})
```

See also:

http://elasticsearch-py.readthedocs.org/en/latest/api.html#elasticsearch.client.IndicesClient.create elasticsearch-py indices.create API documentation

http://www.elasticsearch.org/guide/reference/api/admin-indices-create-index/ Elasticsearch create index API documentation

2.2.4 Types and Mappings

A *type* is a set of fields. A document is of a given type if it has those fields. Whenever you index a document, you specify which type the document is. This is sometimes called a "doctype", "document type" or "doc type".

A *mapping* is the definition of fields and how they should be indexed for a type. In ElasticUtils, we call a document type that has a defined mapping a "mapping type" mostly as a shorthand for "document type with a defined mapping" because that's a mouthful.

Elasticsearch can infer mappings to some degree, but you get a lot more value by specifying mappings explicitly.

To define a mapping, you use *.indices.put_mapping()*.

For example:

```
es = get_es()
es.indices.put_mapping(
    index='blog-index',
    doc_type='blog-entry-type',
    body={
        'blog-entry-type': {
            'properties': {
                'id': {'type': 'integer'},
                'title': {'type': 'string'},
                'content': {'type': 'string'},
                'tags': {'type': 'string'},
                'created': {'type': 'date'}
            }
        }
    }
)
```

You can also define mappings when you create the index:

```
es = get_es()
es.indices.create(
    index='blog-index',
    body={
        'mappings': {
            'blog-entry-type': {
                'properties': {
                    'id': {'type': 'integer'},
                    'title': {'type': 'string'},
                    'content': {'type': 'string'},
                    'tags': {'type': 'string'},
                    'created': {'type': 'date'}
                }
            }
        }
    }
)
```

Note: If there's a possibility of a race condition between creating the index and defining the mapping and some

document getting indexed, then it's good to create the index and define the mappings at the same time.

See also:

http://elasticsearch-py.readthedocs.org/en/latest/api.html#elasticsearch.client.IndicesClient.put_mapping elasticsearch-py indices.put_mapping API documentation

http://www.elasticsearch.org/guide/reference/api/admin-indices-put-mapping/ Elasticsearch put_mapping API documentation

http://www.elasticsearch.org/guide/reference/mapping/ Elasticsearch mapping documentation

2.2.5 Indexing documents

Use *.index()* to index a document.

For example:

```
es = get_es()

entry = {'id': 1,
    'title': 'First post!',
    'content': '<p>First post!</p>',
    'tags': ['status', 'blog'],
    'created': '20130423T16:50:22'
    }

es.index(index='blog-index', doc_type='blog-entry-type', body=entry, id=1)
```

If you're indexing a bunch of documents at the same time, you should use *elasticsearch.helpers.bulk_index()*.

For example:

```
from elasticsearch.helpers import bulk_index

es = get_es()

entries = [{ '_id': 42, ... }, { '_id': 47, ... }]

bulk_index(es, entries, index='blog-index', doc_type='blog-entry-type')
```

See also:

http://elasticsearch-py.readthedocs.org/en/latest/api.html#elasticsearch.Elasticsearch.index elasticsearch-py index API documentation

http://elasticsearch-py.readthedocs.org/en/latest/helpers.html#elasticsearch.helpers.bulk_index elasticsearch-py bulk_index API documentation

http://www.elasticsearch.org/guide/reference/api/index_/ Elasticsearch index API documentation

http://www.elasticsearch.org/guide/reference/api/bulk/ Elasticsearch bulk index API documentation

2.2.6 Deleting documents

You can delete documents with *.delete()*.

For example:

```
es = get_es()

es.delete(index='blog-index', doc_type='blog-entry-type', id=1)
```

See also:

http://elasticsearch-py.readthedocs.org/en/latest/api.html#elasticsearch.Elasticsearch.delete elasticsearch-py delete API documentation

http://www.elasticsearch.org/guide/reference/api/delete/ Elasticsearch delete API documentation

2.2.7 Refreshing

After you index documents, they're not available for searches until after the index is refreshed. By default, the index refreshes every second. If you need the documents to show up in searches before that, call *indices.refresh()*.

For example:

```
es = get_es()

es.indices.refresh(index='blog-index')
```

See also:

http://elasticsearch-py.readthedocs.org/en/latest/api.html#elasticsearch.client.IndicesClient.refresh elasticsearch-py indices.refresh API documentation

http://www.elasticsearch.org/guide/reference/api/admin-indices-refresh/ Elasticsearch refresh API documentation

2.2.8 Delete indexes

You can delete indexes with *.indices.delete()*.

For example:

```
es = get_es()

es.indices.delete(index='blog-index')
```

See also:

http://elasticsearch-py.readthedocs.org/en/latest/api.html#elasticsearch.client.IndicesClient.delete elasticsearch-py indices.delete API documentation

http://www.elasticsearch.org/guide/reference/api/admin-indices-delete-index/ Elasticsearch delete index API documentation

2.2.9 Doing all of this with MappingTypes and Indexables

If you're using MappingTypes, then you can do much of the above using methods and classmethods on `MappingType` and `Indexable` classes. See *Mapping types and Indexables* for more details.

2.3 Mapping types and Indexables

2.3.1 The MappingType class

`elasticutils.MappingType` lets you centralize concerns regarding documents you're storing in your Elasticsearch index.

Lets you tie business logic to search results

When you do searches with MappingTypes, you get back those results as an iterable of MappingTypes by default.

For example, say you had a description field and wanted to have a truncated version of it. You could do it this way:

```
class MyMappingType(MappingType):

    # ... missing code here

    def description_truncated(self):
        return self.description[:100]

results = S(MyMappingType).query(description__text='stormy night')

print list(results)[0].description_truncated()
```

Lets you link source data to search results

You can relate a MappingType to a database model or other source allowing you to link documents in the Elasticsearch index back to their origins in a lazy-loading way. This is done by subclassing MappingType and implementing the `get_object()` method. You can then access the original data using the `object` property.

For example:

```
class MyMappingType(MappingType):

    # ... missing code here

    def get_object(self):
        return self.get_model().objects.get(pk=self._id)

results = S(MyMappingType).filter(height__gte=72)[:1]

first = list(results)[0]

# This prints "height" which comes from the Elasticsearch
# document
print first.height

# This prints "height" which comes from the database data
# that the Elasticsearch document is based on. This is the
# first time ``.object`` is used, so it does the db hit
# here.
print first.object.height
```

DefaultMappingType

The most basic MappingType is the DefaultMappingType which is returned if you don't specify a MappingType and also don't call `elasticutils.S.values_dict()` or `elasticutils.S.values_list()`. The DefaultMappingType lets you access search result fields as instance attributes or as keys:

```
res.description
res['description']
```

The latter syntax is helpful when there are attributes defined on the class that have the same name as the document field or aren't valid Python names.

For more information

See *Types and Mappings* for documentation on defining mappings in the index.

See `elasticutils.MappingType` for documentation on creating MappingTypes.

2.3.2 The Indexable class

`elasticutils.Indexable` is a mixin for `elasticutils.MappingType` that has methods and classmethods for making indexing easier.

2.3.3 Example

Here's an example of a class that subclasses *MappingType* and *Indexable*. It's based on a model called *BlogEntry*.

```
class BlogEntryMappingType(MappingType, Indexable):
    @classmethod
    def get_index(cls):
        return 'blog-index'

    @classmethod
    def get_mapping_type_name(cls):
        return 'blog-entry'

    @classmethod
    def get_model(cls):
        return BlogEntry

    @classmethod
    def get_es(cls):
        return get_es(urls=['http://localhost:9200'])

    @classmethod
    def get_mapping(cls):
        return {
            'properties': {
                'id': {'type': 'integer'},
                'title': {'type': 'string'},
                'tags': {'type': 'string'}
            }
        }

    @classmethod
```

```
    def extract_document(cls, obj_id, obj=None):
        if obj == None:
            obj = cls.get_model().get(id=obj_id)

        doc = {}
        doc['id'] = obj.id
        doc['title'] = obj.title
        doc['tags'] = obj.tags
        return doc

    @classmethod
    def get_indexable(cls):
        return cls.get_model().get_objects()
```

With this, I can write code elsewhere in my project that:

1. gets the mapping type name and mapping for documents of type "blog-entry"
2. gets all the objects that are indexable
3. for each object, extracts the Elasticsearch document data and indexes it

When I create my `elasticutils.S` object, I'd create it like this:

```
s = S(BlogEntryMappingType)
```

and now by default any search results I get back are instances of the *BlogEntryMappingType* class.

2.4 Searching

- *Overview*
- *All about S:* `S`
  - *What is S?*
  - *S is chainable*
  - *S can be typed and untyped*
  - *S can be sliced to return the results you want*
  - *S is lazy*
  - *S results can be returned in many shapes*
- *Where to search*
  - *Specifying connection parameters:* `es`
  - *Specifying indexes to search:* `indexes`
  - *Specifying doctypes to search:* `doctypes`
- *By default, S does a Match All*
- *Queries:* `query`
- *Advanced queries:* `Q` *and* `query_raw`
  - *calling .query() multiple times*
  - *should, must and must_not*
  - *The Q class*
  - *query_raw*
  - *adding new query actions*
- *Filters:* `filter`
- *Advanced filters:* `F` *and* `filter_raw`
  - *and vs. or*
  - *The F class*
  - *filter_raw*
  - *adding new filteractions*
- *Query-time field boosting:* `boost`
- *Ordering:* `order_by`
- *Demoting:* `demote`
- *Highlighting:* `highlight`
- *Suggestions:* `suggest`
- *Facets*
  - *Basic facets:* `facet`
  - *Facet Results*
  - *Facets and scope (filters and global)*
  - *Facets... RAW!:* `facet_raw`
  - *Filter and query facets*
- *Scores and explanations*
  - *Seeing the score*
  - *Getting an explanation:* `explain`

2.4.1 Overview

This chapter covers how to search with ElasticUtils.

2.4.2 All about S: S

What is S?

`elasticutils.S` helps you define an Elasticsearch search.

```
searcher = S()
```

This creates an *untyped* `elasticutils.S` using the defaults:

- uses an `elasticsearch.client.Elasticsearch` instance configured to connect to `localhost` – call `elasticutils.S.es()` to specify connection parameters
- searches across all indexes – call `elasticutils.S.indexes()` to specify indexes
- searches across all doctypes – call `elasticutils.S.doctypes()` to specify doctypes

S is chainable

`elasticutils.S` has methods that return a new S instance with the additional specified criteria. In this way S is chainable and you can reuse S objects for your searches.

For example:

```
s1 = S()

s2 = s1.query(content__text='tabs')

s3 = s2.filter(awesome=True)

s4 = s2.filter(awesome=False)
```

s1, *s2*, and *s3* are all different *S* objects. *s1* is a match all.

s2 has a query.

s3 has everything in *s2* with a `awesome=True` filter.

s4 has everything in *s2* with a `awesome=False` filter.

S can be typed and untyped

When you create an `elasticutils.S` with no type, it's called an *untyped S*. By default, search results for a *untyped S* are returned in the form of a sequence of `elasticutils.DefaultMappingType` instances. You can explicitly state that you want a sequence of dicts or lists, too. See *S results can be returned in many shapes* for more details on how to return results in various formats.

You can also construct a *typed S* which is an *S* with a `elasticutils.MappingType` subclass. By default, search results for a *typed S* are returned in the form of a sequence of instances of that type. See *Mapping types and Indexables* for more about MappingTypes.

S can be sliced to return the results you want

By default Elasticsearch gives you the first 10 results.

If you want something different than that, `elasticutils.S` supports slicing allowing you to get back the specific results you're looking for.

For example:

```
some_s = S()

results = list(some_s)          # returns first 10 results (default)
results = list(some_s[:10])     # returns first 10 results
results = list(some_s[10:20])   # returns results 10 through 19
```

```
```

The slicing is chainable, too:

```
some_s = S()[:10]
first_ten_pitchers = some_s.filter(position='pitcher')
```

Note: The slicing happens on the Elasticsearch side—it doesn't pull all the results back and then slice them in Python. Ew.

Note: Unlike slicing other things in Python, if you choose a start, but no end, then you get 10 results starting with the start.

In other words, this:

```
some_s = S()[10:]
```

does **not** give you all the results from index 10 onwards. Instead it gives you results 10 through 19.

If you want "all the results from index 10 onwards", then you could do something like this:

```
SOME_LARGE_NUMBER = 1000000
some_s = S()[10:SOME_LARGE_NUMBER]
```

If you know you have fewer results than `SOME_LARGE_NUMBER` or you could do this which will kick off two Elasticsearch queries:

```
some_s = S()[10:some_s.count()]
```

Note that doing open-ended queries like this has the same ramifications as calling `elasticutils.S.everything()`. Refer to that documentation for the fearsome details.

See also:

http://www.elasticsearch.org/guide/reference/api/search/from-size.html Elasticsearch from / size documentation

S is lazy

The search won't execute until you do one of the following:

1. use the `elasticutils.S` in an iterable context
2. call `len()` on a `elasticutils.S`
3. call the `elasticutils.S.execute()`, `elasticutils.S.everything()`, `elasticutils.S.count()`, `elasticutils.S.suggestions()` or `elasticutils.S.facet_counts()` methods

Once you execute the search, then it will cache the results and further executions of that `elasticutils.S` won't result in another roundtrip to your Elasticsearch cluster.

S results can be returned in many shapes

An *untyped S* (e.g. `S()`) will return instances of `elasticutils.DefaultMappingType` by default.

A *typed S* (e.g. `S(FooMappingType)`), will return instances of that type (e.g. type `FooMappingType`) by default.

`elasticutils.S.values_list()` gives you a list of tuples. See documentation for more details.

`elasticutils.S.values_dict()` gives you a list of dicts. See documentation for more details.

If you use `elasticutils.S.execute()`, you get back a `elasticutils.SearchResults` instance which has additional useful bits including the raw response from Elasticsearch. See documentation for details.

2.4.3 Where to search

Specifying connection parameters: `es`

`elasticutils.S` will generate an `elasticsearch.client.Elasticsearch` object that connects to `localhost` by default. That's usually not what you want. You can use the `elasticutils.S.es()` method to specify the arguments used to create the elasticsearch-py Elasticsearch object.

Examples:

```
q = S().es(urls=['localhost'])
q = S().es(urls=['localhost:9200'], timeout=10)
```

See `elasticutils.get_es()` for the list of arguments you can pass in.

Specifying indexes to search: `indexes`

An *untyped S* will search all indexes by default.

A *typed S* will search the index returned by the `elasticutils.MappingType.get_index()` method.

If that's not what you want, use the `elasticutils.S.indexes()` method.

For example, this searches all indexes:

```
q = S()
```

This searches just "someindex":

```
q = S().indexes('someindex')
```

This searches "thisindex" and "thatindex":

```
q = S().indexes('thisindex', 'thatindex')
```

This searches whatever `FooMappingType.get_index()` returns:

```
q = S(FooMappingType)
```

Specifying doctypes to search: `doctypes`

An *untyped S* will search all doctypes by default.

A *typed S* will search the doctype returned by the `elasticutils.MappingType.get_mapping_type_name()` method.

If that's not what you want, then you should use the `elasticutils.S.doctypes()` method.

For example, this searches all doctypes:

```
q = S()
```

This searches just the "sometype" doctype:

```
q = S().doctypes('sometype')
```

This searches "thistype" and "thattype":

```
q = S().doctypes('thistype', 'thattype')
```

2.4.4 By default, S does a Match All

By default, `elasticutils.S` with no filters or queries specified will do a `match_all` query in Elasticsearch.

See also:

http://www.elasticsearch.org/guide/reference/query-dsl/match-all-query.html Elasticsearch match_all documentation

2.4.5 Queries: `query`

Queries are specified using the `elasticutils.S.query()` method. See those docs for API details.

ElasticUtils uses this syntax for specifying queries:

fieldname__fieldaction=value

1. fieldname: the field the query applies to
2. fieldaction: the kind of query it is
3. value: the value to query for

The fieldname and fieldaction are separated by __ (that's two underscores).

For example:

```
q = S().query(title__match='taco trucks')
```

will do an Elasticsearch match query on the title field for "taco trucks".

There are many different field actions to choose from:

field action	Elasticsearch query type
(no action specified)	Term query
term	Term query
terms	Terms query
in	Terms query
text	Text query (*DEPRECATED*)
match	Match query [1]
prefix	Prefix query [2]
gt, gte, lt, lte	Range query
range	Range query [3]
fuzzy	Fuzzy query
wildcard	Wildcard query
text_phrase	Text phrase query (*DEPRECATED*)
match_phrase	Match phrase query [1]
query_string	Querystring query [4]
distance	Geo distance query [5]

[1] Elasticsearch 0.19.9 renamed text queries to match queries. If you're using Elasticsearch 0.19.9 or later, you should use match and

See also:

http://www.elasticsearch.org/guide/reference/query-dsl/ Elasticsearch docs for query dsl

http://www.elasticsearch.org/guide/reference/query-dsl/term-query.html Elasticsearch docs on term queries

http://www.elasticsearch.org/guide/reference/query-dsl/terms-query.html Elasticsearch docs on terms queries

http://www.elasticsearch.org/guide/reference/query-dsl/text-query.html Elasticsearch docs on text and text_phrase queries

http://www.elasticsearch.org/guide/reference/query-dsl/match-query.html Elasticsearch docs on match and match_phrase queries

http://www.elasticsearch.org/guide/reference/query-dsl/prefix-query.html Elasticsearch docs on prefix queries

http://www.elasticsearch.org/guide/reference/query-dsl/range-query.html Elasticsearch docs on range queries

http://www.elasticsearch.org/guide/reference/query-dsl/fuzzy-query.html Elasticsearch docs on fuzzy queries

http://www.elasticsearch.org/guide/reference/query-dsl/wildcard-query.html Elasticsearch docs on wildcard queries

http://www.elasticsearch.org/guide/reference/query-dsl/query-string-query.html Elasticsearch docs on query_string queries

http://www.elasticsearch.org/guide/en/elasticsearch/reference/current/query-dsl-geo-distance-filter.html Elasticsearch docs on geo_distance queries

2.4.6 Advanced queries: `Q` and `query_raw`

calling .query() multiple times

Calling `elasticutils.S.query()` multiple times will combine all the queries together.

should, must and must_not

By default all queries must match a document in order for the document to show up in the search results.

You can alter this behavior by flagging your queries with `should`, `must`, and `must_not` flags.

should

A query added with `should=True` affects the score for a result, but it won't prevent the document from being in the result set.

Example:

```
qs = S().query(title__text='castle',
               summary__text='castle',
               should=True)
```

If the document matches either the `title__text` or the `summary__text` then it's included in the results set. It doesn't *have* to match both.

match_phrase. If you're using a version prior to 0.19.9 use text and text_phrase.

[2] You can also use `startswith`, but that's deprecated.

[3] The `range` field action is a shortcut for defining both sides of the range at once. The range is inclusive on both sides and accepts a tuple with the lower value first and upper value second.

[4] When doing `query_string` queries, if the query text is malformed it'll raise a *SearchPhaseExecutionException* exception.

[5] The `distance` field need accepts a tuple with distance, latitude and longitude where distance is a string like 5km.

must

This is the default, so if you don't specify, then it's a *must*.

A query added with `must=True` must match in order for the document to be in the result set.

Example:

```
qs = S().query(title__text='castle',
               summary__text='castle')

qs = S().query(title__text='castle',
               summary__text='castle',
               must=True)
```

These two are equivalent. The document must match both the `title__text` and `summary__text` queries in order to be included in the result set. If it doesn't match one of them, then it's not included.

must_not

A query added with `must_not=True` must NOT match in order for the document to be in the result set.

Example:

```
qs = (S().query(title__text='castle')
         .query(author='castle', must_not=True))
```

For a document to be included in the result set, it must match the `title__text` query and must NOT match the `author` query. I.e. The title must have "castle", but the document can't have been written by someone with "castle" in their name.

The Q class

You can manipulate query units with the `elasticutils.Q` class. For example, you can incrementally build your query:

```
q = Q()

if search_authors:
    q += Q(author_name=search_text, should=True)

if search_keywords:
    q += Q(keyword=search_text, should=True)

q += Q(title__text=search_text, summary__text=search_text,
       should=True)
```

The + Python operator will combine two *Q* instances together and return a new instance.

You can then use one or more *Q* classes in a query call:

```
if search_authors:
    q += Q(author_name=search_text, should=True)

if search_keywords:
    q += Q(keyword=search_text, should=True)

q += Q(title__text=search_text, summary__text=search_text,
       should=True)

s = S().query(q)
```

query_raw

`elasticutils.S.query_raw()` lets you explicitly define the query portion of an Elasticsearch search.

For example:

```
q = S().query_raw({'match': {'title': 'example'}})
```

This will override all `.query()` calls you've made in your `elasticutils.S` before and after the *.query_raw* call.

This is helpful if ElasticUtils is missing functionality you need.

adding new query actions

You can subclass `elasticutils.S` and add handling for additional query actions. This is helpful in two circumstances:

1. ElasticUtils doesn't have support for that query type
2. ElasticUtils doesn't support that query type in a way you need—for example, ElasticUtils uses different argument values

See `elasticutils.S` for more details on how to do this.

2.4.7 Filters: `filter`

Filters are specified using the `elasticutils.S.filter()` method. See those docs for API details.

```
q = S().filter(language='korean')
```

will do a search and only return results where the language is Korean.

`elasticutils.S.filter()` uses the same syntax for specifying fields, actions and values as `elasticutils.S.query()`.

field action	elasticsearch filter
in	Terms filter
gt, gte, lt, lte	Range filter
range	Range filter [6]
prefix, startswith	Prefix filter
(no action)	Term filter

You can also filter on fields that have `None` as a value or have no value:

```
q = S().filter(language=None)
```

This uses the Elasticsearch Missing filter.

Note: In order to filter on fields that have `None` as a value, you have to tell Elasticsearch that the field can have null values. To do this, you have to add `null_value: True` to the mapping for that field.

http://www.elasticsearch.org/guide/reference/mapping/core-types.html

[6] The `range` field action is a shortcut for defining both sides of the range at once. The range is inclusive on both sides and accepts a tuple with the lower value first and upper value second.

See also:

http://www.elasticsearch.org/guide/reference/query-dsl/ Elasticsearch docs for query dsl

http://www.elasticsearch.org/guide/reference/query-dsl/terms-filter.html Elasticsearch docs for terms filter

http://www.elasticsearch.org/guide/reference/query-dsl/range-filter.html Elasticsearch docs for range filter

http://www.elasticsearch.org/guide/reference/query-dsl/prefix-filter.html Elasticsearch docs for prefix filter

http://www.elasticsearch.org/guide/reference/query-dsl/term-filter.html Elasticsearch docs for term filter

http://www.elasticsearch.org/guide/reference/query-dsl/missing-filter.html Elasticsearch docs for missing filter

2.4.8 Advanced filters: `F` and `filter_raw`

and vs. or

Calling filter multiple times is equivalent to an "and"ing of the filters.

For example:

```
q = (S().filter(style='korean')
        .filter(price='FREE'))
```

will do a query for style 'korean' AND price 'FREE'. Anything that has a style other than 'korean' or a price other than 'FREE' is removed from the result set.

You can do the same thing by putting both filters in the same `elasticutils.S.filter()` call.

For example:

```
q = S().filter(style='korean', price='FREE')
```

The F class

Suppose you want either Korean or Mexican food. For that, you need an "or". You can do something like this:

```
q = S().filter(or_={'style': 'korean', 'style':'mexican'})
```

But, wow—that's icky looking and not particularly helpful!

So, we've also got an `elasticutils.F()` class that makes this sort of thing easier.

You can do the previous example with `F` like this:

```
q = S().filter(F(style='korean') | F(style='mexican'))
```

will get you all the search results that are either "korean" or "mexican" style.

What if you want Mexican food, but only if it's FREE, otherwise you want Korean?:

```
q = S().filter(F(style='mexican', price='FREE') | F(style='korean'))
```

F supports `&` (and), `|` (or) and ~ (not) operations.

Additionally, you can create an empty F and build it incrementally:

```
qs = S()
f = F()
if some_crazy_thing:
    f &= F(price='FREE')
if some_other_crazy_thing:
    f |= F(style='mexican')

qs = qs.filter(f)
```

If neither *some_crazy_thing* or *some_other_crazy_thing* are `True`, then F will be empty. That's ok because empty filters are ignored.

filter_raw

`elasticutils.S.filter_raw()` lets you explicitly define the filter portion of an Elasticsearch search.

For example:

```
qs = S().filter_raw({'term': {'title': 'foo'}})
```

This will override all `.filter()` calls you've made in your `elasticutils.S` before and after the *.filter_raw* call.

This is helpful if ElasticUtils is missing functionality you need.

adding new filteractions

You can subclass `elasticutils.S` and add handling for additional filter actions. This is helpful in two circumstances:

1. ElasticUtils doesn't have support for that filter type
2. ElasticUtils doesn't support that filter type in a way you need—for example, ElasticUtils uses different argument values

See `elasticutils.S` for more details on how to do this.

2.4.9 Query-time field boosting: `boost`

ElasticUtils allows you to specify query-time field boosts with `elasticutils.S.boost()`.

These boosts take effect at the time the query is executing. After the query has executed, then the boost is applied and that becomes the final score for the query.

This is a useful way to weight queries for some fields over others.

See `elasticutils.S.boost()` for more details.

Note: Boosts are ignored if you use query_raw.

2.4.10 Ordering: `order_by`

ElasticUtils `elasticutils.S.order_by()` lets you change the order of the search results.

See `elasticutils.S.order_by()` for more details.

See also:

http://www.elasticsearch.org/guide/reference/api/search/sort.html Elasticsearch docs on sort parameter in the Search API

2.4.11 Demoting: `demote`

You can demote documents that match query criteria:

```
q = (S().query(title='trucks')
        .demote(0.5, description__text='gross'))
```

This does a query for trucks, but demotes any that have "gross" in the description with a fraction boost of 0.5.

Note: You can only call `elasticutils.S.demote()` once. Calling it again overwrites previous calls.

This is implemented using the *boosting query* in Elasticsearch. Anything you specify with `elasticutils.S.query()` goes into the *positive* section. The *negative query* and *negative boost* portions are specified as the first and second arguments to `elasticutils.S.demote()`.

Note: Order doesn't matter. So:

```
q = (S().query(title='trucks')
        .demote(0.5, description__text='gross'))
```

does the same thing as:

```
q = (S().demote(0.5, description__text='gross')
        .query(title='trucks'))
```

See also:

http://www.elasticsearch.org/guide/reference/query-dsl/boosting-query.html Elasticsearch docs on boosting query (which are as clear as mud)

2.4.12 Highlighting: `highlight`

ElasticUtils can highlight excerpts for search results.

See `elasticutils.S.highlight()` for more details.

See also:

http://www.elasticsearch.org/guide/reference/api/search/highlighting.html Elasticsearch docs for highlight

2.4.13 Suggestions: `suggest`

Spelling suggestions can be asked for by using the `elasticutils.S.suggest()` method, and then retrieved in `elasticutils.S.suggestions()`:

```
q = S().query(text='Aice').suggest('mysuggest', 'Alice', field='text')
print q.suggestions()['mysuggest'][0]['options']
```

Note: Spelling suggestions require Elasticsearch 0.90 or later.

See also:

http://www.elasticsearch.org/guide/en/elasticsearch/reference/current/search-suggesters.html Elasticsearch docs for suggesters

2.4.14 Facets

Basic facets: `facet`

```
q = (S().query(title='taco trucks')
        .facet('style', 'location'))
```

will do a query for "taco trucks" and return terms facets for the `style` and `location` fields.

Note that the fieldname you provide in the `elasticutils.S.facet()` call becomes the facet name as well.

The facet counts are available through `elasticutils.S.facet_counts()`. For example:

```
q = (S().query(title='taco trucks')
        .facet('style', 'location'))
counts = q.facet_counts()
```

Also, you can get them with the `facets` attribute of the search results:

```
q = (S().query(title='taco trucks')
        .facet('style', 'location'))

results = q.execute()
counts = results.facets
```

You can also restrict the number of terms returned per facet by passing a `size` keyword argument to `elasticutils.S.facet()`:

```
q = S().query(title='taco trucks')
        .facet('style', 'location', size=5)
```

See also:

http://www.elasticsearch.org/guide/reference/api/search/facets/ Elasticsearch docs on facets

http://www.elasticsearch.org/guide/reference/api/search/facets/terms-facet.html Elasticsearch docs on terms facet

Facet Results

The execution methods `elasticutils.S.facet_counts()` and `elasticutils.S.execute()` will return a dictionary containing the named parameter and a `elasticutils.FacetResult` object.

For example:

```
>>> facet_counts = S().facet('primary_country_id').facet_counts()
>>> facet_counts
{u'primary_country_id': <elasticutils.FacetResult at 0x45f12d0>}
```

The FacetResult object contains all of the information returned in the facet stanza.

In the above case, we faceted on primary_country_id as a terms facet. To see the facet results simply iterate over the FacetResult object:

```
>>> for facet_result in facet_counts['primary_country_id']:
...     print facet_result
...
{u'count': 187293, u'term': 41}
    {u'count': 24177, u'term': 9}
    {u'count': 17200, u'term': 50}
    {u'count': 13015, u'term': 15}
    {u'count': 10296, u'term': 30}
    {u'count': 8824, u'term': 32}
    {u'count': 7703, u'term': 6}
    {u'count': 7502, u'term': 23}
    {u'count': 5614, u'term': 2}
    {u'count': 5214, u'term': 33}
```

And to get the "other", "missing" and "total" information from the facetresult:

```
>>> facet_counts['primary_country_id'].missing
3475

>>> facet_counts['primary_country_id'].other
25273

>>> facet_counts['primary_country_id'].total
312111
```

FacetResult is backwords compatible with older versions of ElasticUtils, so you shouldn't need to change anything when upgrading:

```
>>> some_s = S().facet_raw(primary_country_id={'statistical':{"field":"primary_country_id"}})
>>> facet_counts = some_s.facet_counts()
>>> facet_counts['primary_country_id'].max == facet_counts['primary_country_id']['max']
True
```

Facets and scope (filters and global)

What happens if your search includes filters?

Here's an example:

```
q = (S().query(title='taco trucks')
        .filter(style='korean')
        .facet('style', 'location'))
```

The "style" and "location" facets here ONLY apply to the results of the query and are not affected at all by the filters.

If you want your filters to apply to your facets as well, pass in the filtered flag.

For example:

```
q = (S().query(title='taco trucks')
        .filter(style='korean')
        .facet('style', 'location', filtered=True))
```

What if you want the filters to apply just to one of the facets and not the other? You need to add them incrementally.

For example:

```
q = (S().query(title='taco trucks')
        .filter(style='korean')
```

```
        .facet('style', filtered=True)
        .facet('location'))
```

What if you want the facets to apply to the entire corpus and not just the results from the query? Use the *global_* flag.

For example:

```
q = (S().query(title='taco trucks')
        .filter(style='korean')
        .facet('style', 'location', global_=True))
```

Note: The flag name is *global_* with an underscore at the end. Why? Because *global* with no underscore is a Python keyword.

See also:

http://www.elasticsearch.org/guide/reference/api/search/facets/ Elasticsearch docs on facets, facet_filter, and global

http://www.elasticsearch.org/guide/reference/api/search/facets/terms-facet.html Elasticsearch docs on terms facet

Facets... RAW!: `facet_raw`

Elasticsearch facets can do a lot of other things. Because of this, there exists `elasticutils.S.facet_raw()` which will do whatever you need it to. Specify key/value args by facet name.

You could do the first facet example with:

```
q = (S().query(title='taco trucks')
        .facet_raw(style={'terms': {'field': 'style'}}))
```

One of the things this lets you do is scripted facets.

For example:

```
q = (S().query(title='taco trucks')
        .facet_raw(styles={
            'field': 'style',
            'script': 'term == korean ? true : false'
        }))
```

Warning: If for some reason you have specified a facet with the same name using both `elasticutils.S.facet()` and `elasticutils.S.facet_raw()`, the facet_raw stuff will override the facet stuff.

See also:

http://www.elasticsearch.org/guide/reference/modules/scripting.html Elasticsearch docs on scripting

Filter and query facets

You can also define arbitrary facets for queries and facets as documented in Elasticsearch's docs.

For example:

```
q = (S().query(title='taco trucks')
        .facet_raw(korean_or_mexican={
            'filter': {
                'or': [
                    {'term': {'style': 'korean'}},
                    {'term': {'style': 'mexican'}},
                ]
            }
        }))
```

Then access the custom facet via the name you passed into `facet_raw`:

```
counts = q.facet_counts()
korean_or_mexican_count = counts['korean_or_mexican']['count']
```

The same can be done with queries:

```
q = (S().query(title='taco trucks')
      .facet_raw(korean={
          'query': {
              'term': {'style': 'korean'},
          }
      }))
```

See also:

http://www.elasticsearch.org/guide/en/elasticsearch/reference/current/search-facets-query-facet.html
Elasticsearch docs on query facets

http://www.elasticsearch.org/guide/en/elasticsearch/reference/current/search-facets-filter-facet.html
Elasticsearch docs on filter facets

2.4.15 Scores and explanations

Seeing the score

Wondering what the score for a document was? ElasticUtils puts that in the `score` attribute of the `es_meta` object of the search result. For example, let's search an index that holds knowledge base articles for ones with the word "crash" in them and print out the scores:

```
q = S().query(title__text='crash', content__text='crash')

for result in q:
    print result.es_meta.score
```

This works regardless of what form the search results are in.

Getting an explanation: `explain`

Wondering why one document shows up higher in the results than another that should have shown up higher? Wonder how that score was computed? You can set the search to pass the `explain` flag to Elasticsearch with `elasticutils.S.explain()`.

ElasticUtils puts the explanation in the `explanation` attribute of the `es_meta` object of the search result.

For example, let's do a query on a search corpus of knowledge base articles for articles with the word "crash" in them:

```
q = (S().query(title__text='crash', content__text='crash')
        .explain())

for result in q:
    print result.es_meta.explanation
```

This works regardless of what form the search results are in.

See also:

http://www.elasticsearch.org/guide/reference/api/search/explain.html Elasticsearch docs on explain (which are pretty bereft of details).

2.5 More like this: `MLT`

ElasticUtils exposes Elasticsearch More Like This API with the *MLT* class.

For example:

```
mlt = MLT(2034, index='addon_index', doctype='addon')
```

This creates an *MLT* that will return documents that are like document with id 2034 of type *addon* in the *addon_index*.

You can pass it an *S* instance and the *MLT* will derive the index, doctype, ElasticSearch object and also use the search specified by the *S* in the body of the More Like This request. This allows you to get documents like the one specified that also meet query and filter criteria. For example:

```
s = S().filter(product='firefox')
mlt = MLT(2034, s=s)
```

See `elasticutils.MLT` for more details.

See also:

http://www.elasticsearch.org/guide/reference/api/more-like-this.html Elasticsearch guide on More Like This API

http://www.elasticsearch.org/guide/reference/query-dsl/mlt-query.html Elasticsearch guide on the moreLikeThis query which specifies the additional parameters you can use.

http://elasticsearch-py.readthedocs.org/en/latest/api.html#elasticsearch.Elasticsearch.mlt elasticsearch-py documentation for MLT

2.6 Debugging

Here are a few helpful utilities for debugging your ElasticUtils work.

2.6.1 Score explanations

Want to see how a score for a search result was calculated? See *Scores and explanations*.

2.6.2 Logging

elasticsearch-py logs to the `elasticsearch` and `elasticsearch.trace` loggers using the Python logging module.

If you configure `elasticsearch.trace` to show INFO-level messages, then it'll show the requests in curl form, responses if you enable DEBUG.

`elasticsearch` logger will give you information about node failures (WARNING-level), their resurrection (INFO) and every request in a short form (DEBUG). Additionally it will log a WARNING for any failed request.

Elasticsearch-py uses urllib3 by default which logs to the `urllib3` logger using the Python logging module. If you configure that to show INFO-level messages, then you'll see all that stuff. If you configured your elasticsearch-py client to use other transport use it's logging capabilities.

First set up logging using something like this:

```
import logging

# Set up the logging in some way. If you don't have logging
# set up, you can set it up like this.
logging.basicConfig()
```

Then set the logging level for the elasticsearch-py and urllib3 loggers to `logging.DEBUG`:

```
logging.getLogger('elasticsearch').setLevel(logging.DEBUG)
logging.getLogger('urllib3').setLevel(logging.DEBUG)
```

elasticsearch-py will log lines like:

```
INFO:elasticsearch:GET http://localhost:9200/_search [status:200
request:0.001s]
```

Or you can enable the `elasticsearch.trace` logger and have it log a shell transcript of your session using curl:

```
tracer = logging.getLogger('elasticsearch.trace')
tracer.setLevel(logging.DEBUG)
tracer.addHandler(logging.FileHandler('/tmp/elasticsearch-py.sh'))
```

Note: The trace logger will always point to localhost:9200 and add `?pretty` to the query string of the url so that when you're curling, then Elasticsearch will return a prettified response that's easier to read.

2.6.3 Getting the search body

The *S* class has a *build_search()* method that you can use to see the body of the Elasticsearch search request it generates with the parameters you've specified. This is helpful in debugging ElasticUtils and figuring out whether it's doing things poorly.

For example:

```
some_s = S()
print some_s.build_search()
```

We also have `elasticutils.utils.to_json()` which takes the output of `elasticutils.S.build_search()` and returns the JSON string. This is helpful if you need to take the search body that ElasticUtils generates and tinker with it using curl or elasticsearch-head.

2.6.4 elasticsearch-head

https://github.com/mobz/elasticsearch-head

elasticsearch-head is the phpmyadmin for elasticsearch. It makes it much easier to see what's going on.

2.6.5 elasticsearch-paramedic

https://github.com/karmi/elasticsearch-paramedic

elasticsearch-paramedic allows you to see the state and real-time statistics of your Elasticsearch cluster.

2.6.6 es2unix

https://github.com/elasticsearch/es2unix

Use this for calling Elasticsearch API things instead of curl.

2.7 API docs

- *Functions*
- *The S class*
- *The F class*
- *The Q class*
- *The SearchResults class*
- *The MappingType class*
- *The Indexable class*
- *The DefaultMappingType class*
- *The MLT class*
- *The ESTestCase class*
- *Helper utilites*

2.7.1 Functions

2.7.2 The S class

2.7.3 The F class

2.7.4 The Q class

2.7.5 The SearchResults class

2.7.6 The MappingType class

2.7.7 The Indexable class

2.7.8 The DefaultMappingType class

2.7.9 The MLT class

2.7.10 The ESTestCase class

2.7.11 Helper utilites

2.8 Migrating from Elasticsearch 0.90 to 1.x with ElasticUtils

Note: This is a work in progress and probably doesn't cover everything.

2.8.1 Summary

There are a bunch of API-breaking changes between Elasticsearch 0.90 and 1.x. Because of this, it's really tricky to get over this hump without having downtime.

This document covers a high-level walk through for upgrading from Elasticsearch 0.90 to 1.x and the steps you should take to reduce your downtime.

Note: "1.x" covers 1.0, 1.1 and 1.2.

2.8.2 Steps

Each of these steps should result in a working system. Do them one at a time and test everything in between.

1. Upgrade to ElasticUtils 0.9.1

   You must use elasticsearch-py version 0.4.5–don't use a later version!

2. Upgrade your Elasticsearch cluster to version 0.90.13

3. Upgrade to ElasticUtils 0.10.1

   You will need to update elasticsearch-py past 0.4.5. The latest version should work fine.

4. Make any changes to your code so that it works with both Elasticsearch 0.90 and 1.x

   There are some tricky things here, see the *Tricky things* section.

5. Upgrade to Elasticsearch 1.x

At that point, you should be using a recent version of the elasticsearch-py library and a recent version of Elasticsearch and should be all set.

2.8.3 Resources

See also:

http://www.elasticsearch.org/guide/en/elasticsearch/reference/current/breaking-changes.html Breaking changes when migrating to Elasticsearch 1.0

http://www.elasticsearch.org/guide/en/elasticsearch/reference/current/_deprecations.html Deprecated features when migrating to Elasticsearch 1.0

2.8.4 Tricky things

There are a few tricky differences between Elasticsearch 0.90 and 1.0 that will affect your code.

Changes with `.values_dict()` and `.values_list()`

Explanation

In Elasticsearch 1.x, you get back different shapes of things depending on whether you specify "fields". To smooth this out and normalize the differences between Elasticsearch 0.90 and 1.x, ElasticUtils now **always** passes in `fields` property when you use `elasticutils.S.values_list()` and `elasticutils.S.values_dict()`.

Let's show some code to illustrate the new behavior.

First, a bunch of setup:

```
>>> from elasticutils import get_es, S
>>> from elasticsearch.helpers import bulk_index
>>> URL = 'localhost'
>>> INDEX = 'fooindex'
>>> BOOK_DOCTYPE = 'book'
>>> PERSON_DOCTYPE = 'person'
>>> es = get_es(urls=[URL])
>>> es.indices.delete(index=INDEX, ignore=404)
```

Now define the two document mappings we're going to use: book and person. Book has no stored fields. Person has two.

```
>>> mapping = {
...     BOOK_DOCTYPE: {
...         'properties': {
...             'id': {'type': 'integer'},
...             'title': {'type': 'string'},
...             'tags': {'type': 'string'},
...             }
...         },
...     PERSON_DOCTYPE: {
...         'properties': {
```

```
...             'id': {'type': 'integer', 'store': True},
...             'name': {'type': 'string', 'store': True},
...             'weight': {'type': 'integer'}
...         }
...     }
... }
```

Create the index with the mappings, add some books and add some people.

```
>>> es.indices.create(INDEX, body={'mappings': mapping})
>>> books = [
...     {'_id': 1, 'id': 1, 'title': '10 Balloons', 'tags': ['kids', 'hardcover']},
...     {'_id': 2, 'id': 2, 'title': 'Puppies', 'tags': ['animals']},
...     {'_id': 3, 'id': 3, 'title': 'Dictionary', 'tags': ['reference']},
... ]
>>> bulk_index(es, books, index=INDEX, doc_type=BOOK_DOCTYPE)
(3, [])
>>> people = [
...     {'_id': 1, 'id': 1, 'name': 'Bob', 'weight': 40},
...     {'_id': 2, 'id': 2, 'name': 'Jim', 'weight': 44},
...     {'_id': 3, 'id': 3, 'name': 'Jim Bob', 'weight': 42},
... ]
>>> bulk_index(es, people, index=INDEX, doc_type=PERSON_DOCTYPE)
[...]
>>> es.indices.refresh(index=INDEX)
[...]
```

Now let's do some queries so we can see how things work now.

Let's build a `basic_s` that looks at our Elasticsearch cluster and the index. Also a `book_s` and a `person_s`.

```
>>> basic_s = S().es(urls=[URL]).indexes(INDEX)
>>> book_s = basic_s.doctypes(BOOK_DOCTYPE)
>>> person_s = basic_s.doctypes(PERSON_DOCTYPE)
```

How many documents are in our index?

```
>>> basic_s.count()
6
```

Call `.values_list()` on books which has no stored fields so we get back the `_id` and `_type` for each document returned and all values are lists:

```
>>> list(book_s.values_list())
[([u'1'], [u'book']), ([u'2'], [u'book']), ([u'3'], [u'book'])]
```

`.values_list('id')` on books, so we get id returned and all values are lists:

```
>>> list(book_s.values_list('id'))
[([1],), ([2],), ([3],)]
```

`.values_list()` on persons which does have stored fields (id and name, but not weight), so we get the stored fields returned and all values are lists:

```
>>> list(person_s.values_list())
[([1], [u'Bob']), ([2], [u'Jim']), ([3], [u'Jim Bob'])]
```

`.values_list('id')` on persons which works just like books because we've specified which fields we want back:

```
>>> list(person_s.values_list('id'))
[([1],), ([2],), ([3],)]
```

The same goes for `.values_dict()`.

What you need to do

1. If you have calls to `.values_list()` and `.values_dict()` that don't specify any fields, then you either need to change the mapping and store the fields you want back, or change the calls so they specify the fields you want back.
2. Every time you use results from a `.values_list()` or `.values_dict()` call, you need to change it to always treat the values as lists.

CHAPTER 3

Using ElasticUtils with Django

3.1 Using ElasticUtils with Django

- *Summary*
- *How to integrate ElasticUtils with Django*
- *Configuration*
- *Elasticsearch*
- *Using with Django ORM models*
- *Celery tasks*
- *Middleware*
- *Writing tests*
- *Helpful things to know*
  - *Indexing and reset_queries*

3.1.1 Summary

Django-specific code is all located in *elasticutils.contrib.django*.

This chapter covers using ElasticUtils Django bits. For API documentation, see *Django API docs*.

3.1.2 How to integrate ElasticUtils with Django

1. add ElasticUtils configuration settings to your project's setting file
2. write one or more *MappingType* classes
3. write code to create the Elasticsearch index and populate it with documents based on your *MappingType* subclasses
3. use `elasticutils.contrib.django.S` to search and return results
4. use `elasticutils.contrib.django.estestcase.ESTestCase` to write tests

That's the gist of it. You can deviate on any of these depending on your needs, of course.

3.1.3 Configuration

ElasticUtils depends on the following settings in your Django settings file:

django.conf.settings.**ES_DISABLED**
If *ES_DISABLED* = *True*, then Any method wrapped with *es_required* will return and log a warning. This is useful while developing, so you don't have to have Elasticsearch running.

django.conf.settings.**ES_URLS**
This is a list of Elasticsearch urls. In development this will look like:

```
ES_URLS = ['http://localhost:9200']
```

django.conf.settings.**ES_INDEXES**
This is a mapping of doctypes to indexes. A *default* mapping is required for types that don't have a specific index.

When ElasticUtils queries the index for a model, by default it derives the doctype from *Model._meta.db_table*. When you build your indexes and mapping types, make sure to match the indexes and mapping types you're using.

Example 1:

```
ES_INDEXES = {'default': 'main_index'}
```

This only has a default, so all ElasticUtils queries will look in *main_index* for all mapping types.

Example 2:

```
ES_INDEXES = {'default': 'main_index',
              'splugs': 'splugs_index'}
```

Assuming you have a *Splug* model which has a *Splug._meta.db_table* value of *splugs*, then ElasticUtils will run queries for *Splug* in the *splugs_index*. ElasticUtils will run queries for other models in *main_index* because that's the default.

Example 3:

```
ES_INDEXES = {'default': ['main_index'],
              'splugs': ['splugs_index']}
```

FIXME: The API allows for this. Pretty sure it should query multiple indexes, but we have no tests for that and I haven't tested it, either.

django.conf.settings.**ES_TIMEOUT**
Default: 5

The timeout in seconds for creating the Elasticsearch connection.

3.1.4 Elasticsearch

The *get_es()* in the Django contrib will use Django settings listed above to build the elasticsearch-py Elasticsearch object.

3.1.5 Using with Django ORM models

Requirements Django

The *elasticutils.contrib.django.S* class takes a *MappingType* in the constructor. That allows you to tie Django ORM models to Elasticsearch index search results.

In `elasticutils.contrib.django` is *MappingType* which has some additional Django ORM-specific code in it to make it easier.

Define a *MappingType* subclass for your model. The minimal you need to define is *get_model*.

Further, you can use the *Indexable* mixin to get a bunch of helpful indexing-related code.

For example, here's a minimal *MappingType* subclass:

```
from django.models import Model
from elasticutils.contrib.django import MappingType


class MyModel(Model):
    # Django model ...


class MyMappingType(MappingType):
    @classmethod
    def get_model(cls):
        return MyModel

searcher = MyMappingType.search()
```

Here's one that uses *Indexable* and handles indexing:

```
from django.models import Model
from elasticutils.contrib.django import Indexable, MappingType


class MyModel(Model):
    # Django model ...


class MyMappingType(MappingType, Indexable):
    @classmethod
    def get_model(cls):
        """Returns the Django model this MappingType relates to"""
        return MyModel

    @classmethod
    def get_mapping(cls):
        """Returns an Elasticsearch mapping for this MappingType"""
        return {
            'properties': {
                # The id is an integer, so store it as such. Elasticsearch
                # would have inferred this just fine.
                'id': {'type': 'integer'},

                # The name is a name---so we shouldn't analyze it
                # (de-stem, tokenize, parse, etc).
                'name': {'type': 'string', 'index': 'not_analyzed'},

                # The bio has free-form text in it, so analyze it with
                # snowball.
                'bio': {'type': 'string', 'analyzer': 'snowball'},

                # Age is an integer
                'age': {'type': 'integer'}
            }
        }

    @classmethod
```

```
    def extract_document(cls, obj_id, obj=None):
        """Converts this instance into an Elasticsearch document"""
        if obj is None:
            obj = cls.get_model().objects.get(pk=obj_id)

        return {
            'id': obj.id,
            'name': obj.name,
            'bio': obj.bio,
            'age': obj.age
            }


searcher = MyMappingType.search()
```

See also:

http://www.elasticsearch.org/guide/reference/mapping/ The Elasticsearch guide on mapping types.

http://www.elasticsearch.org/guide/reference/mapping/core-types.html The Elasticsearch guide on mapping type field types.

3.1.6 Celery tasks

Requirements Django, Celery

You can then utilize things such as `elasticutils.contrib.django.tasks.index_objects()` to automatically index all new items.

3.1.7 Middleware

Requirements Django

There's a middleware that catches all Elasticsearch-related exceptions and shows a 501/503 template accordingly. See `elasticutils.contrib.django.ESExceptionMiddleware` for details.

3.1.8 Writing tests

Requirements Django

When writing test cases for your ElasticUtils-using code, you'll want to do a few things:

1. Default `ES_DISABLED` to *True*. This way, the tests that kick off creating data but aren't testing search-specific things don't additionally index stuff. That'll save you a bunch of test time.
2. When testing ElasticUtils things, override the settings and set `ES_DISABLED` to *False*.
3. Use an `ESTestCase` that sets up the indexes before tests run and tears them down after they run.
4. When testing, make sure you use an index name that's unique. You don't want to run your tests and have them affect your production index.

You can use `elasticutils.contrib.django.estestcase.ESTestCase` for your app's tests. It's pretty basic but does all of the above except item 1 which you'll need to do in your test settings.

Example usage:

```
from elasticutils.contrib.django.estestcase import ESTestCase


class TestQueries(ESTestCase):
    # This class holds tests that do elasticsearch things

    def test_query(self):
        # Test code ...

    def test_locked_filters(self):
        # Test code ...
```

ElasticUtils uses this for it's Django tests. Look at the test code for more examples of usage:

https://github.com/mozilla/elasticutils/

If it's not what you want, you could subclass it and override behavior or just write your own.

3.1.9 Helpful things to know

Indexing and reset_queries

If you are:

1. indexing a lot of data pulled out with the Django ORM, and
2. have `DEBUG = True` (i.e. development environments)

then you'll probably want to call `django.db.reset_queries()` periodically.

What's going on is that when `DEBUG = True` (i.e. a devleopment environment), Django helpfully stores all the queries that are made which when you're indexing a lot of data is a lot of data. Calling `django.db.reset_queries()` periodically flushes the queries so it doesn't monotonically eat all your memory before the indexing is done.

3.2 Django API docs

- *The S class*
- *The MappingType class*
- *The Indexable class*
- *View decorators*
- *The ESExceptionMiddleware class*
- *Tasks*
- *The ESTestCase class*

3.2.1 The S class

3.2.2 The MappingType class

3.2.3 The Indexable class

3.2.4 View decorators

3.2.5 The ESExceptionMiddleware class

3.2.6 Tasks

3.2.7 The ESTestCase class

Subclass this and make it do what you need it to do. It's definitely worth reading the code.

CHAPTER 4

Contributor's Guide

4.1 Join this project!

Interested in working on a Python library for using elasticsearch? Interested in using it? Then you should be interested in this project!

4.1.1 Want to help?

Here are things we need help with:

- **fixing bugs listed in the issue tracker**
- **writing tests**
- **writing documentation**: We could use help writing better documentation for ElasticUtils.
- **spreading the word**: Do you know other people who would like this software? If so, tell them about ElasticUtils!
- **project infrastructure**: Is there infrastructure that's missing in this project that would make it easier for you to collaborate? If so, what?

Are you thinking, "That list is makes me want to go shopping for bumper stickers!" That's ok! Hop on IRC, say hi and we can go from there!

For project details, see *ElasticUtils*.

4.2 Hacking HOWTO

This covers setting up a development environment for developing on ElasticUtils. If you're interested in using ElasticUtils, then you should check out *User's Guide*.

4.2.1 External requirements

You should have Elasticsearch installed and running.

4.2.2 Install dependencies

Run:

```
$ virtualenv ./venv
$ . ./venv/bin/activate
$ pip install -r requirements/dev.txt
$ python setup.py develop
```

This sets up the required dependencies for development of ElasticUtils.

Note: You don't have to put your virtual environment in `./venv/`. Feel free to put it anywhere.

4.3 Conventions

We follow the code conventions listed in the coding conventions page of the webdev bootcamp guide. This covers all the Python code.

We use git and follow the conventions listed in the git and github conventions page of the webdev bootcamp guide.

4.4 Documentation

4.4.1 Conventions

See the docmentation page in the webdev bootcamp guide for documentation conventions.

The documentation is available in HTML and PDF forms at http://elasticutils.readthedocs.org/. This tracks documentation in the master branch of the git repository. Because of this, it is always up to date.

4.4.2 Building the docs

The documentation in *docs/* is built with Sphinx. To build HTML version of the documentation, do:

```
$ cd docs/
$ make html
```

4.5 Running and writing tests

4.5.1 Running the tests

You can run the tests with:

```
./run_tests.py
```

This will run all the tests.

Note: If you need to adjust the settings, copy `test_settings.py` to a new file (like `test_settings_local.py`), edit the file, and specify that as the value for the environment variable `DJANGO_SETTINGS_MODULE`.

DJANGO_SETTINGS_MODULE=test_settings_local ./run_tests.py

This is helpful if you need to change the value of `ES_HOSTS` to match the ip address or port that elasticsearch is listening on.

4.5.2 Writing tests

Tests are located in *elasticutils/tests/*.

We use nose for test utilities and running tests.

4.5.3 ElasticTestCase

If you're testing things in ElasticUtils that require hitting an Elasticsearch cluster, then you should subclass *elasticutils.tests.ESTestCase* which has code in it for making things easier.

4.6 Release process

1. Checkout master tip.
2. Check to make sure `setup.py`, requirements files, and `docs/installation.rst` have correct version of elasticsearch-py.
3. Update version numbers in `elasticutils/_version.py`.
   (a) Set `__version__` to something like `0.4`.
   (b) Set `__releasedate__` to something like `20120731`.
4. Update `CONTRIBUTORS`, `CHANGELOG`, `MANIFEST.in`.

   Make sure to set the date for the release in CHANGELOG.

   Make sure requirements in `setup.py`, `docs/installation.rst` and `CHANGELOG` all match.
5. Verify correctness.
   (a) Run tests.
   (b) Build docs.
   (c) Run sample programs in docs.
   (d) Verify all that works.
6. Tag the release:

```
$ git tag -a v0.4
```

   Copy the details from `CHANGELOG` into the tag comment.

7. Push everything:

```
$ git push --tags official master
```

8. Update PyPI:

```
$ rm -rf dist/*
$ python setup.py sdist bdist_wheel
$ twine upload dist/*
```

9. Update topic in `#elasticutils`, blog post, twitter, etc.

CHAPTER 5

Sample programs

5.1 Basic sample program

Here's a short script that gives you the gist of how to use ElasticUtils:

```
"""
This is a sample program that uses Elasticsearch (from elasticsearch-py)
object to create an index, create a mapping, and index some data. Then
it uses ElasticUtils S to show some behavior.
"""

from elasticutils import get_es, S

from elasticsearch.helpers import bulk_index

URL = 'localhost'
INDEX = 'fooindex'
DOCTYPE = 'testdoc'


# This creates an elasticsearch.Elasticsearch object which we can use
# to do all our indexing.
es = get_es(urls=[URL])

# First, delete the index if it exists.
es.indices.delete(index=INDEX, ignore=404)

# Define the mapping for the doctype 'testdoc'. It's got an id field,
# a title which is analyzed, and two fields that are lists of tags, so
# we don't want to analyze them.
mapping = {
    DOCTYPE: {
        'properties': {
            'id': {'type': 'integer'},
            'title': {'type': 'string', 'analyzer': 'snowball'},
            'topics': {'type': 'string'},
            'product': {'type': 'string', 'index': 'not_analyzed'},
            }
        }
    }

# Create the index 'testdoc' mapping.
es.indices.create(INDEX, body={'mappings': mapping})
```

```
39
40
41  # Let's index some documents and make them available for searching.
42  documents = [
43      {'_id': 1,
44       'title': 'Deleting cookies',
45       'topics': ['cookies', 'privacy'],
46       'product': ['Firefox', 'Firefox for mobile']},
47      {'_id': 2,
48       'title': 'What is a cookie?',
49       'topics': ['cookies', 'privacy'],
50       'product': ['Firefox', 'Firefox for mobile']},
51      {'_id': 3,
52       'title': 'Websites say cookies are blocked - Unblock them',
53       'topics': ['cookies', 'privacy', 'websites'],
54       'product': ['Firefox', 'Firefox for mobile', 'Boot2Gecko']},
55      {'_id': 4,
56       'title': 'Awesome Bar',
57       'topics': ['tips', 'search', 'user interface'],
58       'product': ['Firefox']},
59      {'_id': 5,
60       'title': 'Flash',
61       'topics': ['flash'],
62       'product': ['Firefox']}
63      ]
64
65  bulk_index(es, documents, index=INDEX, doc_type=DOCTYPE)
66  es.indices.refresh(index=INDEX)
67
68
69  # Now let's do some basic queries.
70
71  # Let's build a basic S that looks at our Elasticsearch cluster and
72  # the index and doctype we just indexed our documents in.
73  basic_s = S().es(urls=[URL]).indexes(INDEX).doctypes(DOCTYPE)
74
75  # How many documents are in our index?
76  print basic_s.count()
77  # Prints:
78  # 5
79
80  # Print articles with 'cookie' in the title.
81  print [item['title']
82         for item in basic_s.query(title__match='cookie')]
83  # Prints:
84  # [u'Deleting cookies', u'What is a cookie?',
85  #  u'Websites say cookies are blocked - Unblock them']
86
87  # Print articles with 'cookie' in the title that are related to
88  # websites.
89  print [item['title']
90         for item in basic_s.query(title__match='cookie')
91                            .filter(topics='websites')]
92  # Prints:
93  # [u'Websites say cookies are blocked - Unblock them']
94
95  # Print articles in the 'search' topic.
96  print [item['title']
```

```
       for item in basic_s.filter(topics='search')]
# Prints:
# [u'Awesome Bar']

# Do a query and use the highlighter to denote the matching text.
print [(item['title'], item.es_meta.highlight['title'])
       for item in basic_s.query(title__match='cookie').highlight('title')]
# Prints:
# [
#     (u'Deleting cookies', [u'Deleting <em>cookies</em>']),
#     (u'What is a cookie?', [u'What is a <em>cookie</em>?']),
#     (u'Websites say cookies are blocked - Unblock them',
#        [u'Websites say <em>cookies</em> are blocked - Unblock them']
#     )
# ]


# That's the gist of it!
```

5.2 Sample program using facets

```
"""
This is a sample program that uses Elasticsearch (from elasticsearch-py)
object to create an index, create a mapping, and index some data. Then
it uses ElasticUtils S to show some behavior with facets.
"""

from elasticutils import get_es, S

from elasticsearch.helpers import bulk_index

URL = 'localhost'
INDEX = 'fooindex'
DOCTYPE = 'testdoc'


# This creates an elasticsearch.Elasticsearch object which we can use
# to do all our indexing.
es = get_es(urls=[URL])

# First, delete the index, ignore possible 404 - it means the index doesn't
# exist, so there's nothing to delete.
es.indices.delete(index=INDEX, ignore=404)

# Define the mapping for the doctype 'testdoc'. It's got an id field,
# a title which is analyzed, and two fields that are lists of tags, so
# we don't want to analyze them.
#
# Note: The alternative for the tags is to analyze them and use the
# 'keyword' analyzer. Both not analyzing and using the keyword
# analyzer treats the values as a single term rather than tokenizing
# them and treating as multiple terms.
mapping = {
    DOCTYPE: {
        'properties': {
            'id': {'type': 'integer'},
```

```
            'title': {'type': 'string'},
            'topics': {'type': 'string'},
            'product': {'type': 'string', 'index': 'not_analyzed'},
            }
        }
    }

# create the index with defined mappings
es.indices.create(index=INDEX, body={'mappings': mapping})


# This indexes a series of documents each is a Python dict.
documents = [
    {'_id': 1,
     'title': 'Deleting cookies',
     'topics': ['cookies', 'privacy'],
     'product': ['Firefox', 'Firefox for mobile']},
    {'_id': 2,
     'title': 'What is a cookie?',
     'topics': ['cookies', 'privacy', 'basic'],
     'product': ['Firefox', 'Firefox for mobile']},
    {'_id': 3,
     'title': 'Websites say cookies are blocked - Unblock them',
     'topics': ['cookies', 'privacy', 'websites'],
     'product': ['Firefox', 'Firefox for mobile', 'Boot2Gecko']},
    {'_id': 4,
     'title': 'Awesome Bar',
     'topics': ['tips', 'search', 'basic', 'user interface'],
     'product': ['Firefox']},
    {'_id': 5,
     'title': 'Flash',
     'topics': ['flash'],
     'product': ['Firefox']}
    ]

bulk_index(es, documents, index=INDEX, doc_type=DOCTYPE)

# Elasticsearch will refresh the indexes and make those documents
# available for querying in a second or so (it's configurable in
# Elasticsearch), but we want them available right now, so we refresh
# the index.
es.indices.refresh(index=INDEX)

# Let's build a basic S that looks at the right Elasticsearch cluster,
# index and doctype.
basic_s = S().es(urls=[URL]).indexes(INDEX).doctypes(DOCTYPE).values_dict()

# Now let's see facet counts for all the products.
s = basic_s.facet('product')

print s.facet_counts()
# Pretty-printed output:
# {u'product': {
#     u'_type': u'terms',
#     u'total': 9,
#     u'terms': [
#         {u'count': 5, u'term': u'Firefox'},
#         {u'count': 3, u'term': u'Firefox for mobile'},
```

```
#         {u'count': 1, u'term': u'Boot2Gecko'}
#     ],
#     u'other': 0,
#     u'missing': 0
#     }}

# Let's do a query for 'cookie' and do a facet count.
print s.query(title__match='cookie').facet_counts()
# Pretty-printed output:
# {u'product': {
#     u'_type': u'terms',
#     u'total': 2,
#     u'terms': [
#         {u'count': 1, u'term': u'Firefox for mobile'},
#         {u'count': 1, u'term': u'Firefox'}
#     ],
#     u'other': 0,
#     u'missing': 0
#     }}

# Note that the facet_counts are affected by the query.

# Let's do a filter for 'flash' in the topic.
print s.filter(topics='flash').facet_counts()
# Pretty-printed output:
# {u'product': {
#     u'_type': u'terms',
#     u'total': 9,
#     u'terms': [
#         {u'count': 5, u'term': u'Firefox'},
#         {u'count': 3, u'term': u'Firefox for mobile'},
#         {u'count': 1, u'term': u'Boot2Gecko'}
#     ],
#     u'other': 0,
#     u'missing': 0
#     }}

# Note that the facet_counts are NOT affected by filters.

# Let's do a filter for 'flash' in the topic, and specify
# filtered=True.
print s.facet('product', filtered=True).filter(topics='flash').facet_counts()
# Pretty-printed output:
# {u'product': {
#     u'_type': u'terms',
#     u'total': 1,
#     u'terms': [
#         {u'count': 1, u'term': u'Firefox'}
#     ],
#     u'other': 0,
#     u'missing': 0
#     }}

# Using filtered=True causes the facet_counts to be affected by the
# filters.

# We've done a bunch of faceting on a field that is not
# analyzed. Let's look at what happens when we try to use facets on a
```

```
# field that is analyzed.
print basic_s.facet('topics').facet_counts()
# Pretty-printed output:
# {u'topics': {
#     u'_type': u'terms',
#     u'total': 14,
#     u'terms': [
#         {u'count': 3, u'term': u'privacy'},
#         {u'count': 3, u'term': u'cookies'},
#         {u'count': 2, u'term': u'basic'},
#         {u'count': 1, u'term': u'websites'},
#         {u'count': 1, u'term': u'user'},
#         {u'count': 1, u'term': u'tips'},
#         {u'count': 1, u'term': u'search'},
#         {u'count': 1, u'term': u'interface'},
#         {u'count': 1, u'term': u'flash'}
#     ],
#     u'other': 0,
#     u'missing': 0
#     }}

# Note how the facet counts shows 'user' and 'interface' as two
# separate terms even though they're a single topic for document with
# id=4. When that document is indexed, the topic field is analyzed and
# the default analyzer tokenizes it splitting it into two terms.
#
# Moral of the story is that you want fields you facet on to be
# analyzed as keyword fields or not analyzed at all.
```

CHAPTER 6

Indices and tables

- genindex

d

django.conf.settings, 47

D

django.conf.settings (module), 47

E

ES_DISABLED (in module django.conf.settings), 47
ES_INDEXES (in module django.conf.settings), 48
ES_TIMEOUT (in module django.conf.settings), 48
ES_URLS (in module django.conf.settings), 48