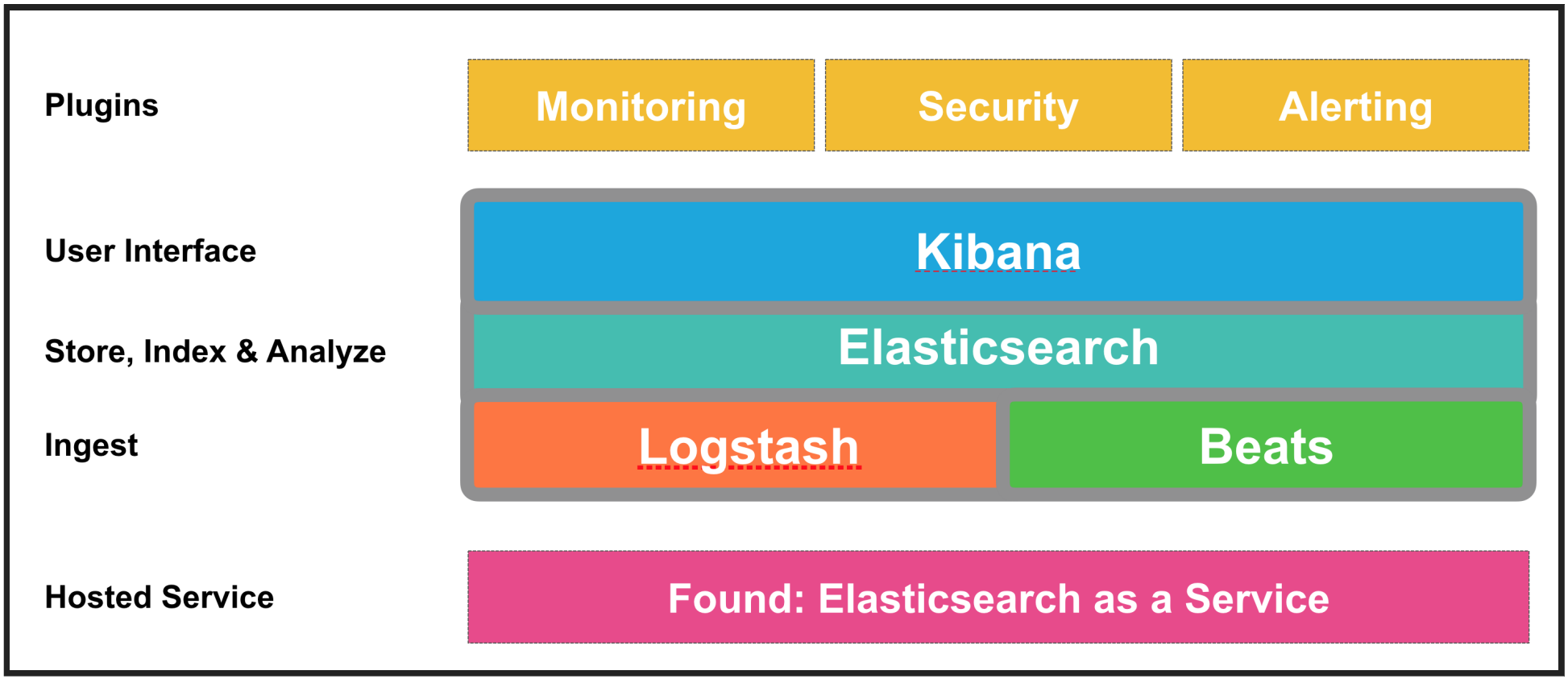
WHAT'S NEW IN ELASTICSEARCH ~~2.0~~2.2 (AND 2.3?)

David Pilato / @dadoonet

ABOUT ME

- Developer | Evangelist at elastic
- Joined: Jan 2013
- Elasticsearch user since February 2011 - 0.14

ABOUT ELASTIC



ELASTICSEARCH 2.0

Very large release

>2,500 Pull Requests

469 commiters

FOCUS ON GROWING UP

FOCUS ON GROWING UP

- Simplification
- Security
- Resiliency
- Features
- Plugins

SIMPLIFICATION

SOMETIMES INCLUDES REMOVING STUFF

WHAT'S NOT IN ELASTICSEARCH 2.0

- Rivers
  - Logstash - *Indexing Twitter With Logstash and Elasticsearch*
  - Own ingestion layer - *Advanced Search for Your Legacy Application*
- Facets - *replaced by aggregations*
- _shutdown API - *use platform specific services*
- Support for Thrift and Memcached protocols
- Bulk UDP - *use the standard bulk API, or use UDP to send documents to Logstash first.*

WHAT'S NOT IN CORE (MOVED TO A PLUGIN)

- Delete by query
- Types:
  - murmur3
  - size
- Multicast discovery

SIMPLIFICATION: MAPPING CHANGES

SIMPLIFICATION: MAPPING CHANGES

- Conflicting field mappings
- Fields cannot be referenced by short name
- Type name prefix removed
- Field names cannot contain dots
- Type names cannot start with a dot
- Type may no longer be deleted
- `index_analyzer` is removed
- `_analyzer` field is removed
- date format changes
- ... and more ...

CONFLICTING FIELD MAPPING

```
PUT my_index
{
  "mappings": {
    "type_one": {
      "properties": {
        "name": { "type": "string" }                          ❶
      }
    },
    "type_two": {
      "properties": {
        "name": { "type": "string", "analyzer": "english" }   ❷
      }
    }
  }
}
```

AMBIGUOUS FIELD LOOKUP BEFORE 2.0

```
PUT my_index
{
  "mappings": {
    "name": {                                  ❶
      "properties": {
        "title": { "type": "string" },         ❷
        "name": {                              ❸
          "properties": {
            "title": { "type": "string" }      ❹
          }
        }
      }
    }
  }
}
```

What does `title` refer to?

What about `name.title`?

What about `name.name.title`?

FIELD LOOKUP IN 2.0

```
PUT my_index
{
  "mappings": {
    "name": {                                   ❶
      "properties": {
        "title": { "type": "string" },          ❷
        "name": {                               ❸
          "properties": {
            "title": { "type": "string" }       ❹
          }
        }
      }
    }
  }
}
```

"`title`" always refers to ❷

"`name.title`" always refers to ❹

"`name.name.title`" is invalid

FIELD ANALYZER

```
PUT my_index
{
  "mappings": {
    "my_type": {
      "properties": {
        "title": { "type": "string", "analyzer": "my_analyzer }
      }
    }
  }
}
```

FIELD ANALYZER

Before 2.0:

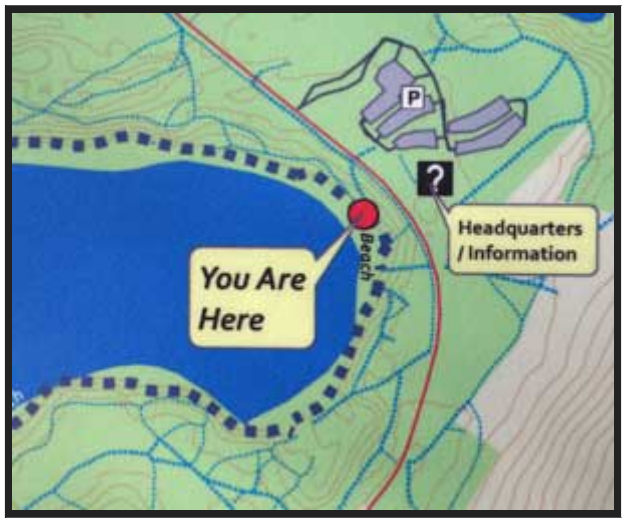
- `analyzer` - sets **index** and **search** analyzers
- `search_analyzer` - sets **search** analyzer
- `index_analyzer` - sets **index** analyzer

Starting with 2.0:

- `analyzer` - sets **index** and **search** analyzers
- `search_analyzer` - overrides **search** analyzer

OTHER MAPPING CHANGES

Check The Great Mapping Refactoring blog post



SIMPLIFICATION: QUERY AND FILTER EXECUTION CHANGES

BEFORE 2.0

Queries:

- Typically contribute to scoring
- No caching

Filters:

- Don't contribute to scoring
- Can be cached

STARTING WITH 2.0

- Filters and queries are merged into queries
- The behavior of a query clause depends on whether it is used in query context or in filter context

BEFORE 2.0

```
{
  "filtered" : {
    "query": { query definition },
    "filter": { filter definition }
  }
}
```

STARTING WITH 2.0

```
{
  "bool" : {
    "must": { query definition },
    "filter": { filter definition }
  }
}
```

TWO-PHASE EXECUTION

Approximation phase

quickly iterates over a superset of the matching documents

Verification phase

check if a document in this superset actually matches the query

TWO-PHASE EXECUTION - HOW IS THIS USEFUL?

```
{
  "bool" : {
    "must": [{
      "match_phrase": {
        "body": "quick fox"  ❶
      }, {
      "match_phrase": {
        "body": "brown dog"  ❷
      }
    }]
  }
}
```

Instead of loading posting list for all documents that contain quick and fox or brown and dog, we only load postings for the documents that contain all 4 terms.

QUERY CACHING

- Fully automatic
- Keeps track of 256 most recently used queries
- Only caches those that appear 5 times or more
- Does not cache segments which have less than 10000 documents or 3% of the documents of the index
- More efficient query cache (roaring bitmaps)
- Non-scoring components are cache-able

SECURITY

NETWORKING CHANGES

Elasticsearch now binds to local interfaces

Unicast discovery is now default

Makes elasticsearch more secure by default

RUNNING UNDER SECURITY MANAGER BY DEFAULT

- Prevents outside access outside of elasticsearch even if elasticsearch process is compromised
- All resources that elasticsearch can access are defined on node startup
- Some libs/plugins are unsecured!

RELIABILITY

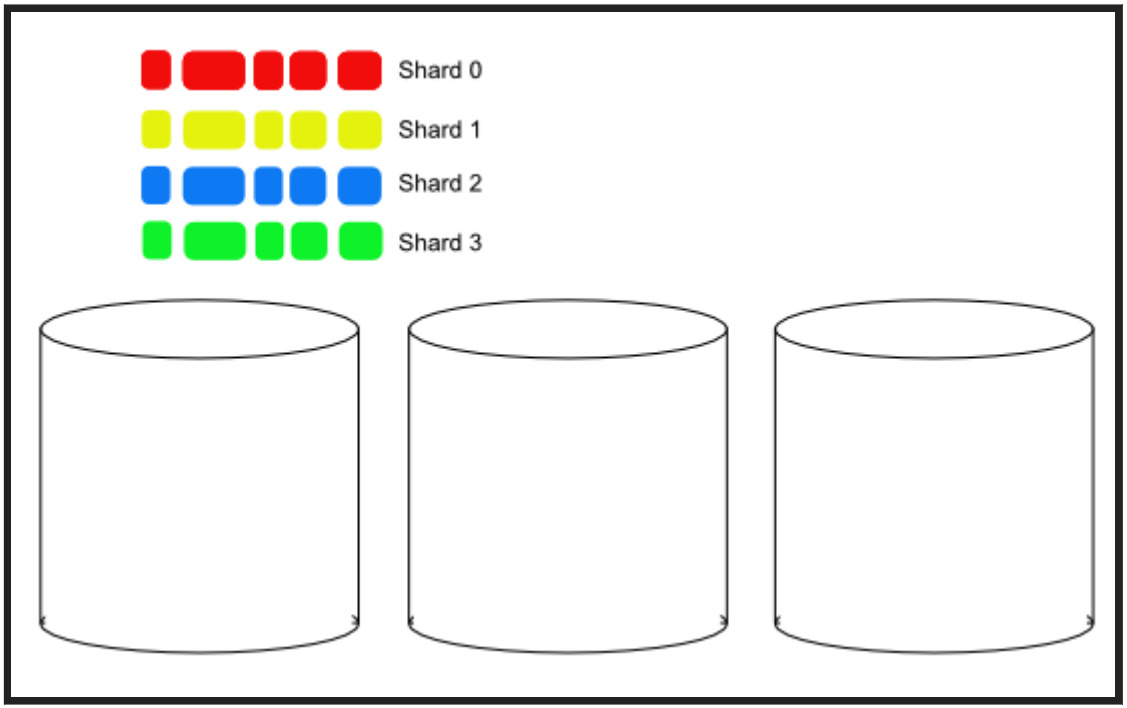
INDEX OPERATIONS ARE NOW DURABLE BY DEFAULT

- Before 2.0 transaction log was fsynced every 5 sec
- Transaction log is now fsync after each operation
- Configurable
- On SSDs have indexing is about 7% - 10% slower with bulk indexing compared to async translog flushes

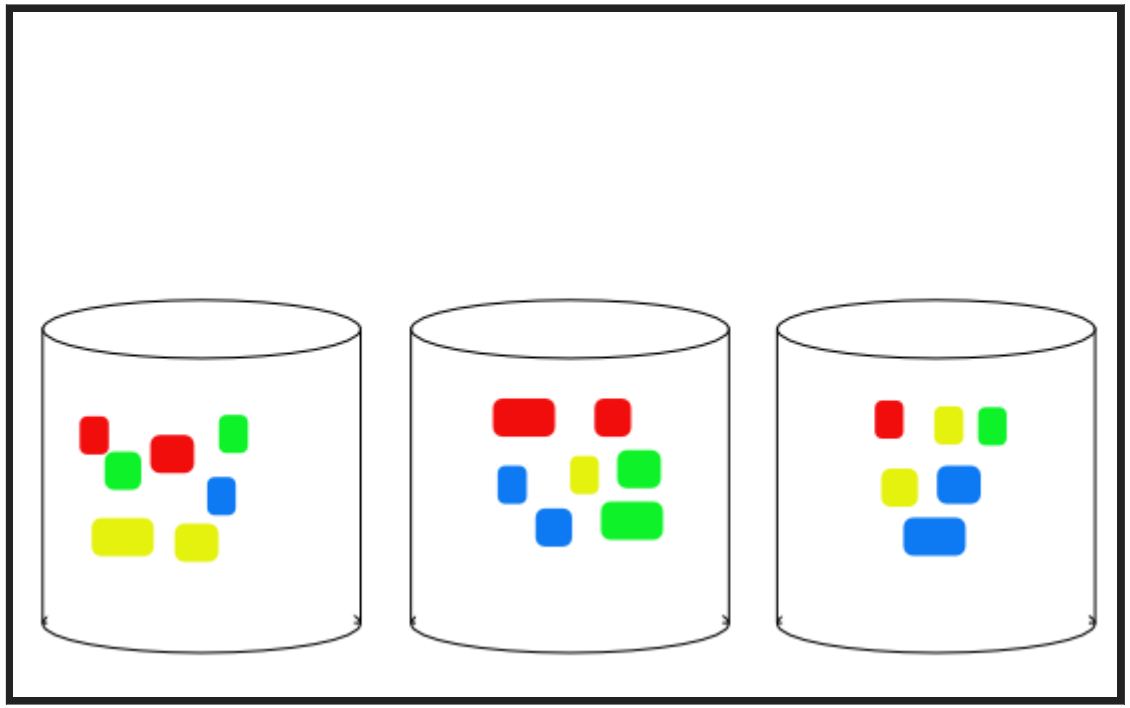
MULTIPLE DATA PATH STRIPING

- Before all shards were stripped across all paths
- This striping is no longer supported. Instead, different shards may be allocated to different paths, but all of the files in a single shard will be written to the same path.

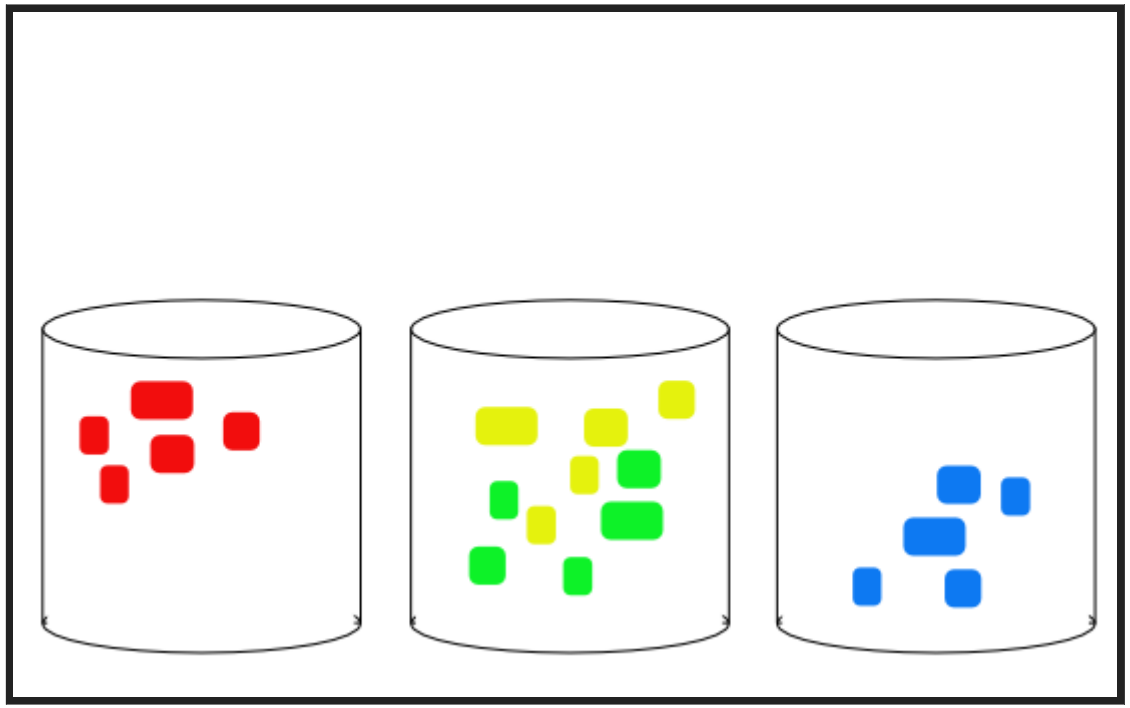
MULTIPLE DATA PATH STRIPING



BEFORE 2.0



AFTER 2.0



CLUSTER STATE DIFFS

- Before entire cluster state was shipped on every change to every node
- Starting with 2.0 only changes are sent

UNITS ARE REQUIRED IN ALL SETTINGS

```
curl -XPUT "localhost:9200/test/_settings" -d '{
    "index" : {
        "refresh_interval" : "5"
    }
}'
```

What have I just done here?

DOC VALUES BY DEFAULT

- Dramatic memory-reduction by default
- Any field that is indexed but not analyzed now has doc values.
- Only for indices created with 2.0

SOME COOL RECENT RELIABILITY CHANGES THAT YOU MIGHT HAVE MISSED

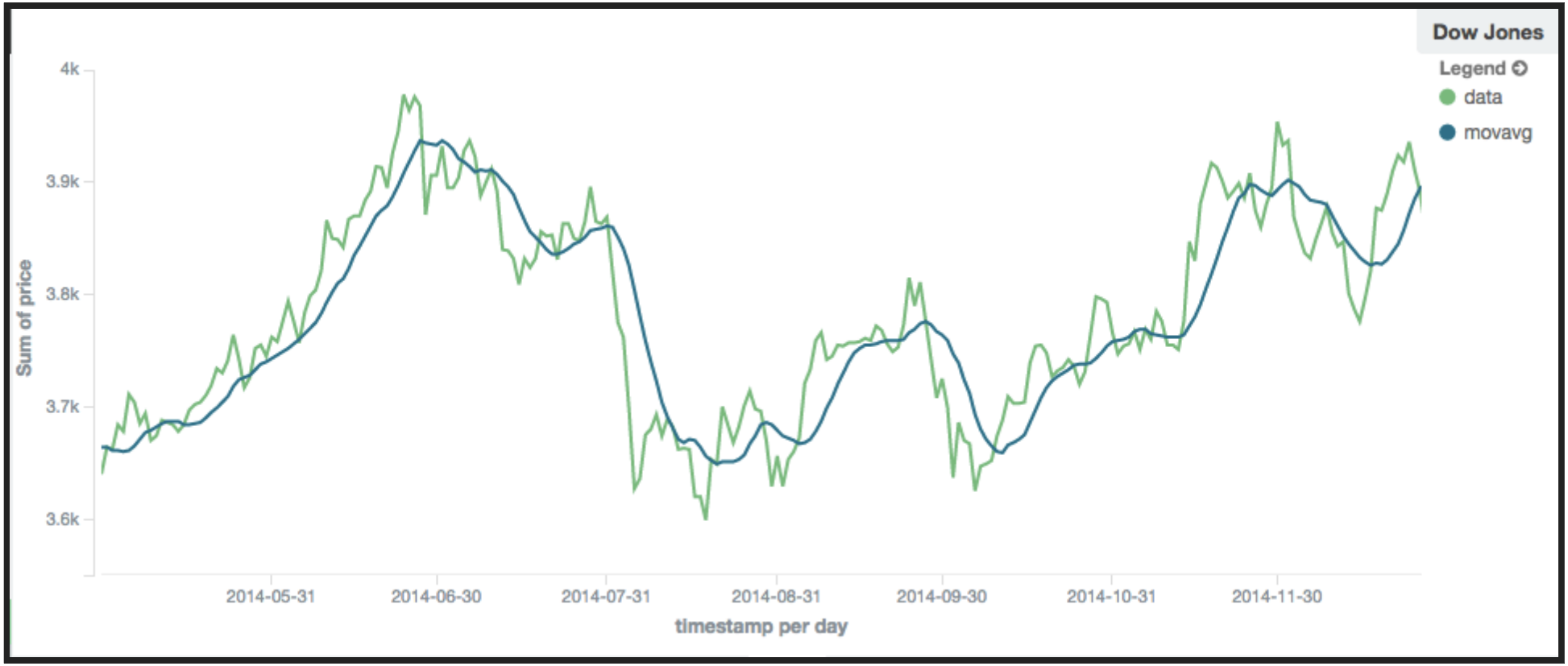
- Sync-flush with `sync_id` (1.6)
- Async shard allocation (1.6) *was blocking cluster state*
- Better handling of node leave/rejoin (1.7) *wait for one minute*

FEATURES

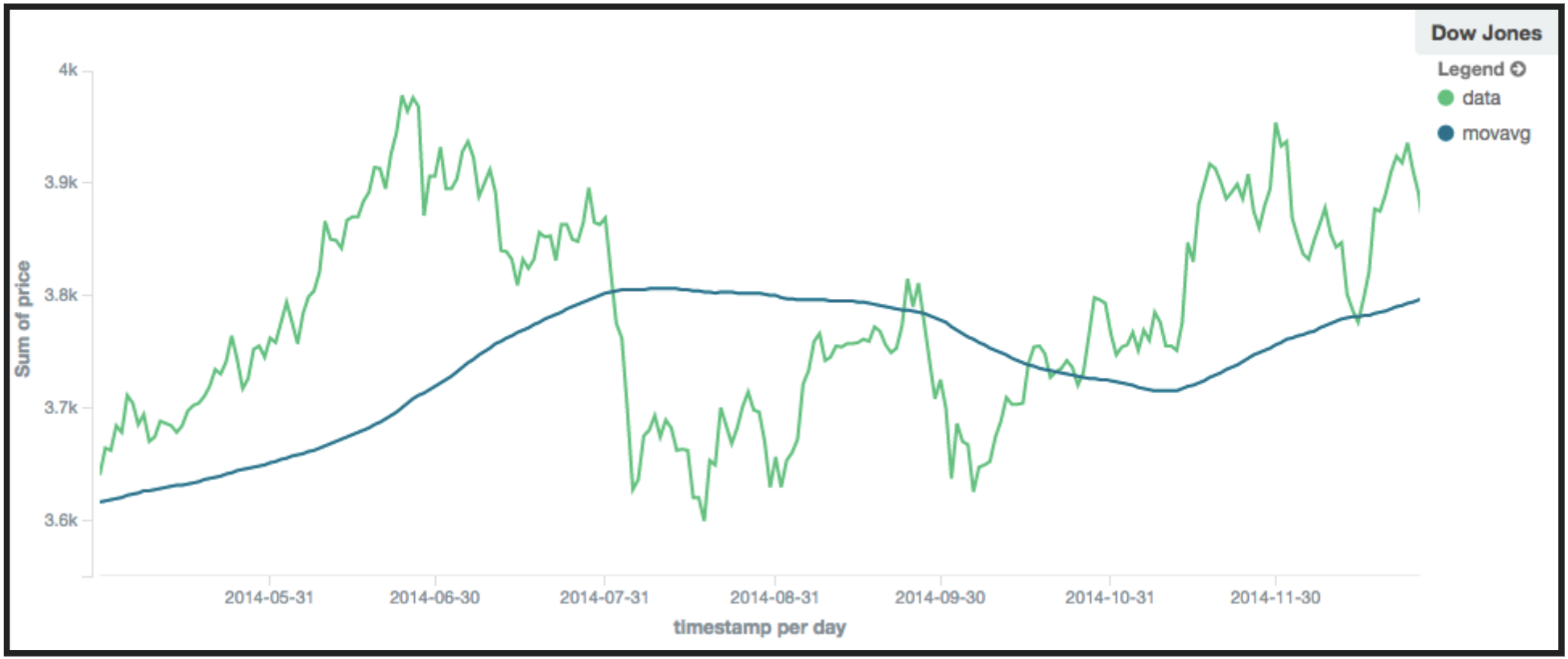
PIPELINE AGGS

- Derivatives
- Moving average
- Holt Winters (prediction / anomaly detection)
- Stats: Min/Max/avg
- Time-series math

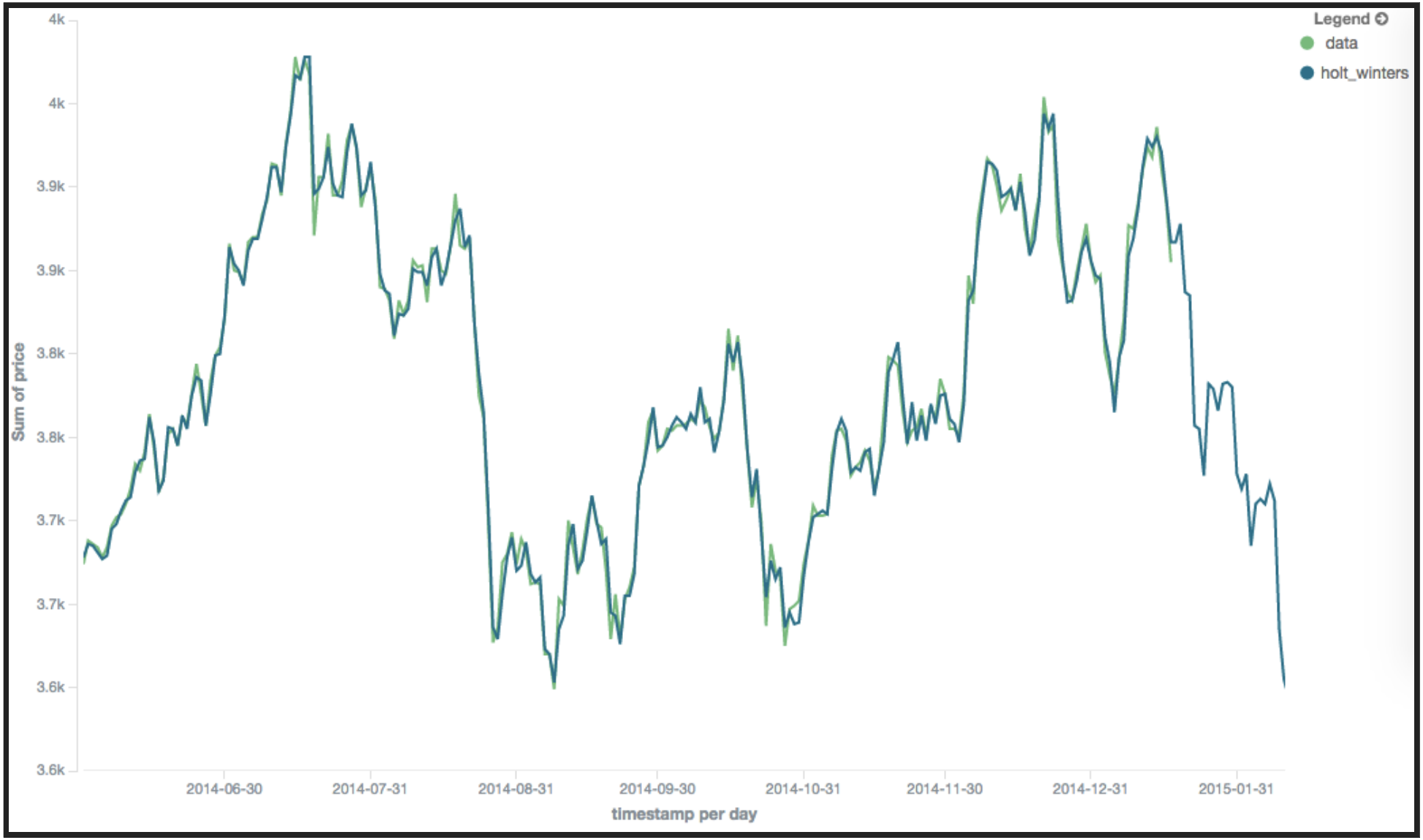
MOVING AVERAGE - 10



MOVING AVERAGE - 100



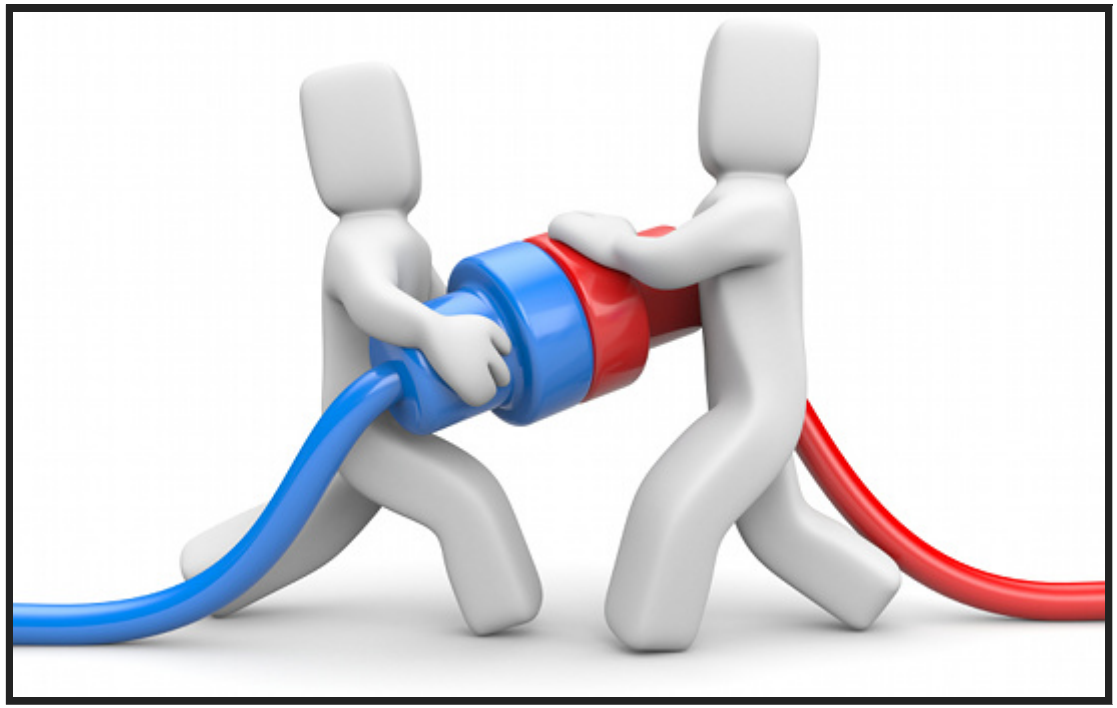
HOLT WINTERS - SEASONAL FLUCTUATIONS



INDEX COMPRESSION

- 10-30% reduction in index size
- Some indexing/merging impact
- Dynamic setting - could be set before optimization for time-based indices

PLUGINS



```
$ mvn clean install

[INFO] Plugin: Analysis: ICU ............................. SUCCESS
[INFO] Plugin: Analysis: Japanese (kuromoji) .............. SUCCESS
[INFO] Plugin: Analysis: Phonetic ......................... SUCCESS
[INFO] Plugin: Analysis: Smart Chinese (smartcn) .......... SUCCESS
[INFO] Plugin: Analysis: Polish (stempel) ................. SUCCESS
[INFO] Plugin: Delete By Query ............................ SUCCESS
[INFO] Plugin: Discovery: Azure ........................... SUCCESS
[INFO] Plugin: Discovery: EC2 ............................. SUCCESS
[INFO] Plugin: Discovery: Google Compute Engine ........... SUCCESS
[INFO] Plugin: Discovery: Multicast ....................... SUCCESS
[INFO] Plugin: Language: Expression ....................... SUCCESS
[INFO] Plugin: Language: Groovy ........................... SUCCESS
[INFO] Plugin: Language: JavaScript ....................... SUCCESS
[INFO] Plugin: Language: Python ........................... SUCCESS
[INFO] Plugin: Mapper: Murmur3 ............................ SUCCESS
[INFO] Plugin: Mapper: Size ............................... SUCCESS
```

BEFORE 2.0

```
$ bin/plugin install elasticsearch/cloud-aws/2.7.2
$ bin/plugin install cloud-aws --url http://p.to/cloud-aws-2.7.2.zip
$ bin/plugin install cloud-aws --url file:p/to/cloud-aws-2.7.2.zip
```

STARTING WITH 2.0

```
$ bin/plugin install cloud-aws
$ bin/plugin install http://p.to/cloud-aws-2.7.2.zip
$ bin/plugin install file:p/to/cloud-aws-2.7.2.zip
```

ELASTICSEARCH 2.1

Geo-centroid aggregation

New pipeline aggregations: stats_bucket, extended_stats_bucket and percentiles_bucket

Allocation and recovery enhancements

Page size max defaults to 10 000

Update only if _source changed

Deprecate count API: size=0

Deprecate scan API: sort by _doc

ELASTICSEARCH 2.2

Query profiler

```
curl -XGET 'localhost:9200/_search' -d '{
  "profile": true,
  "query" : {
    "match" : { "message" : "search test" }
  }
}
```

Improved geo-point fields

Stricter security for plugins and scripting

COMING IN ELASTICSEARCH 2.3

Task Management API

Reindex API

...

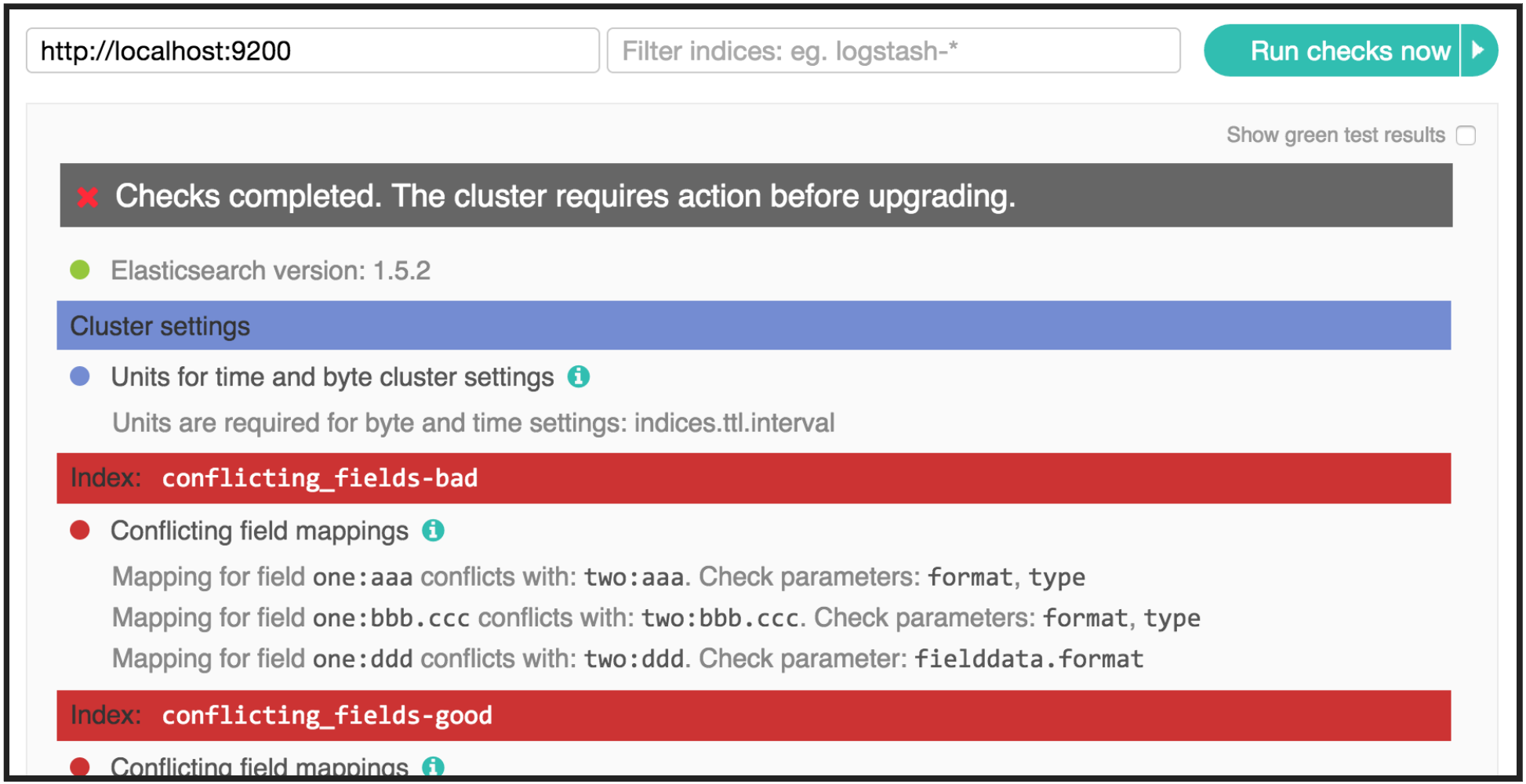
MIGRATION PROCESS

2.0 IS A MAJOR RELEASE

- No rolling upgrades
- One way - no way to downgrade back to 1.x
- Make sure you have a backup
- Test it! Don't try to "wing it" in production.

MIGRATION PLUGIN

- Site plugin for 1.x that checks for potential issues
- https://github.com/elastic/elasticsearch-migration



GETTING HELP

- https://discuss.elastic.co/ - Discussion Forums
- https://github.com/elastic/elasticsearch - Report issues
- IRC on Freenode - #elasticsearch
- http://stackoverflow.com/questions/tagged/elasticsearch

ELASTIC{ON} '16

- Feb 17-19, 2016
- Pier 48, San Francisco, CA
- https://www.elastic.co/elasticon

elastic
user group

- Slides: http://david.pilato.fr/presentations/
- Twitter: @dadoonet