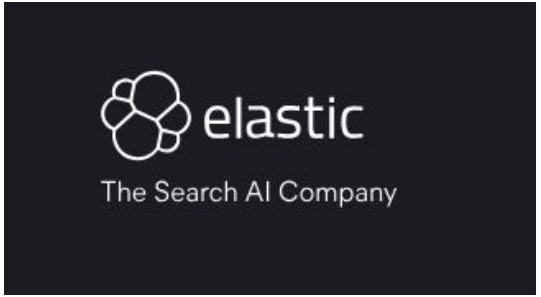
Elasticsearch

Science fair presentasjon

Anders Petershagen Åsbø & Ingebrigt Hovind

Litt bakgrund



- Elasticsearch utvikles av Elastic B. V. (amerikansk-nederlandsk)
- Open source søkemotor og analyseverktøy
- Eget skjema-løst databasesystem basert på json-dokumenter
- Abonnement-basert skyløsning som Elastic NV hoster
- Egne avtaler for self-hosting av Elasticsearch hvis ønsket

TRUSTED BY 50% OF THE FORTUNE 500 TO DRIVE INNOVATION

cisco SITECORE COMCAST DISCOVER Booking.com

Brand Video - Sept 2024. (n.d.). [Video]. Elastic.
https://www.elastic.co/about/

Bruksområder

- Håndtere store mengder data for systematisering og analyse
  - Feilsøking over stort antall program-logger fra ett eller flere programmer
  - Analysere kunders kjøpshistorikk.
- Håndtere store databaser med høy spørring-trafikk
  - Produktsøk i netthandel.
  - Søking i sosiale medier.
- Brukes for eksempel av Wells Fargo for å holde styr på finansielle transaksjoner i (nesten) sanntid
- Er optimalisert for høyt output fra store databaser.
- Vi skal fokusere på hvordan skaleringen skjer og hvordan dataen blir distribuert

Wells Fargo deploys Elastic Observability for distributed tracing and APM. (n.d.). Elastic Customers. https://www.elastic.co/customers/wells-fargo

Hvordan skalerer Elasticsearch?

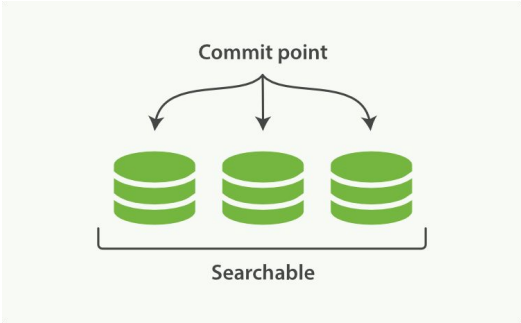
- For å virkelig skalere, så må man skalere horisontalt
  - Men man vil abstrahere dette bort fra brukeren
  - Elasticsearch er distribuert av natur
- En spørring utføres derfor mot én indeks
  - Indeks er en logisk inndeling av data
  - Men indeksen peker “under panseret” mot ett eller flere **shards**
    - Hvert shard er bygd på Apache Lucene, og inneholder én eller flere inverterte indekser

```
PUT my-index-000001/_doc/2?refresh=true
{
  "text": "Document with ID 2"
}
```

_id field | Elasticsearch Guide [8.16] | Elastic. (n.d.). Elastic. https://www.elastic.co/guide/en/elasticsearch/reference/current/mapping-id-field.html

Dynamically updatable indices | *ElasticSearch: The Definitive Guide [2.X]* | *Elastic*. (n.d.). Elastic. https://www.elastic.co/guide/en/elasticsearch/guide/2.x/dynamic-indices.html

Spørringer mot Shards



- Elasticsearch vet ikke hvilket Shard som vil kunne svare på en spørring
- Spørringen deles opp i en query-fase og en fetch-fase
  - I query-fasen så spørres hvert shard, og det returnerer topp-N dokumenter
    - Disse blir merget til én felles rangert liste over dokumenter
  - I fetch-fasen så hentes de faktiske dokumentene, fra listen over

Distributed Search Execution | *ElasticSearch: The Definitive Guide [2.X]* | *Elastic*. (n.d.). Elastic. https://www.elastic.co/guide/en/elasticsearch/guide/2.x/distributed-search.html

Skriving til Shards

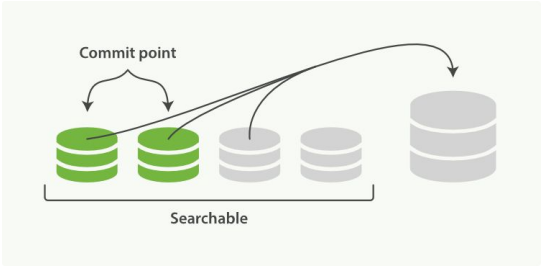
```
PUT /website/blog/123
{
  "title": "My first blog entry",
  "text":  "Just trying this out...",
  "date":  "2014/01/01"
}
```

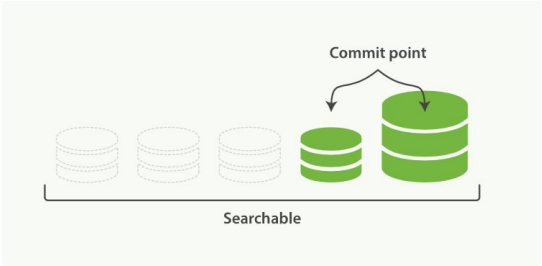
- For å finne hvilket Shard som skal holde hvilket dokument så brukes hash-verdien av ID-en til dokumentet
  - `shard = hash(routing) % number_of_primary_shards`
- Når man skriver til et dokument så oppdateres primary-Shardet først.
  - Dersom dette går bra så oppdateres alle replicas
    - Forespørselen returneres ikke før alle Replicas har blitt oppdatert
  - Gjøres først i cache, og så til disk.

Dynamically updatable indices | ElasticSearch: The Definitive Guide [2.X] | Elastic. (n.d.). Elastic. https://www.elastic.co/guide/en/elasticsearch/guide/2.x/dynamic-indices.html
Routing a document to a shard | ElasticSearch: The Definitive Guide [2.X] | Elastic. (n.d.). Elastic. https://www.elastic.co/guide/en/elasticsearch/guide/2.x/routing-value.html

Hvordan håndtere endring eller fjerning av data?

- Hver shard vinner ytelse på å ha immutable inverterte indekser.
  - Men hvordan kan vi da gjøre endringer i databasen?
- Soft-delete:
  - Marker dokument som slettet uten å faktisk røre dokumentet.
  - Hvis endret: Lag oppdatert kopi av dokumentet i en ny invertert indeks.
- Spørringer vil nå ignorere utdatert data.
- Obs!
  - Minnebruk vokser raskt.
  - Kjør clean-up sykler ved gjevne mellomrom.
    - Rekonstruer indeksene i hver shard med bare levende dokumenter.
    - Slett gamle indekser.
    - Gjøres i bakgrunnen.





Making changes persistent | ElasticSearch: The Definitive Guide [2.X] | Elastic. (n.d.). Elastic. https://www.elastic.co/guide/en/elasticsearch/guide/2.x/translog.html
Segment Merging | ElasticSearch: The Definitive Guide [2.X] | Elastic. (n.d.). Elastic. https://www.elastic.co/guide/en/elasticsearch/guide/2.x/merge-process.html

Oppsummering

- Elasticsearch parallelliserer spørringer over flere forskjellige noder
  - Hver node inneholder ett eller flere shards
    - Og hvert shard inneholder en del av en invertert indeks
- Unngår å måtte tilpasse applikasjonen til spørringen
  - Kan dermed late som om det kun er en eneste stor database