How to setup a search engine

Francesco Piccinno

stack.box@gmail.com
piccinno@di.unipi.it
http://github.com/nopper

Micro demos:
https://github.com/nopper/if2013-search-engine/

Outline

Outline

- Backend
  - Crawling (Scrapy)
  - Data storage (Cassandra)
  - Post processing (Spark)
  - Indexing (Elasticsearch)

Outline

- Backend
  - Crawling (Scrapy)
  - Data storage (Cassandra)
  - Post processing (Spark)
  - Indexing (Elasticsearch)
- Frontend = query module
  - User experience
  - Satisfy or even predict the user needs

Crawling: Scrapy

Why?

- Easy to setup and to play with
- Create a Scrapy project
- Define the fields you are interested in each downloaded file
- Create one spider per set of fields, and assign it a list of sites (static!)
- Define zero or more (*item*-)pipelines for post-processing and data storage
- Plug Redis to be distributed

Demo

Data storage: Cassandra

Why?

- Scalable
- Write performance
- Distributed
- Easy to setup and to play with
- Easy to understand data model

(NoSQL) Data model

(NoSQL) Data model

row ➡ com.google.www:80:/

(NoSQL) Data model

row ➡ com.google.www:80:/

title ➡ “Google”

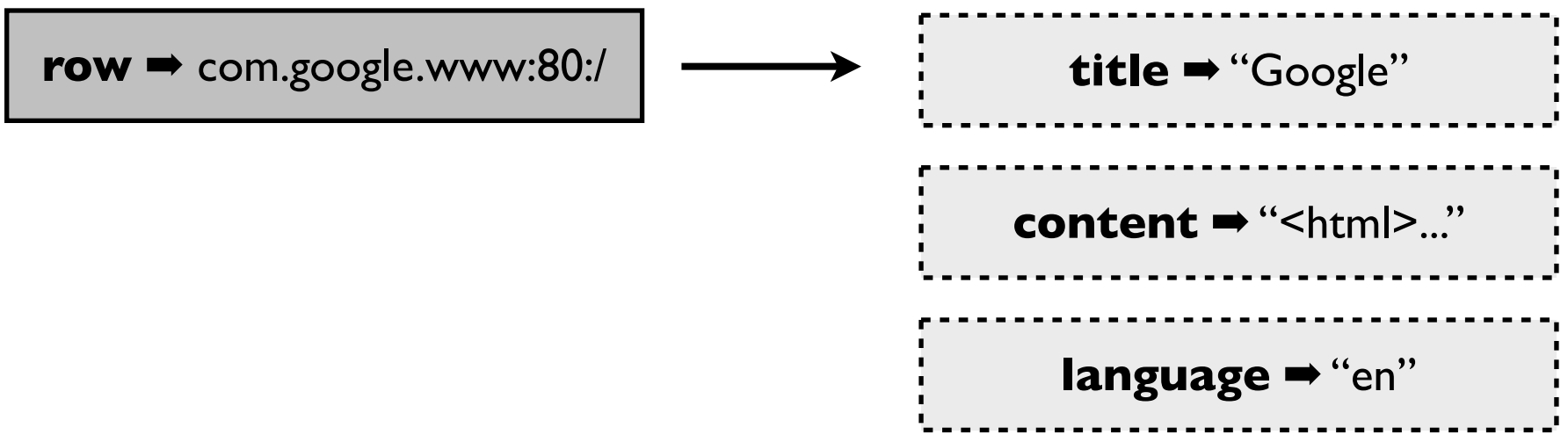
(NoSQL) Data model

row ➡ com.google.www:80:/

title ➡ “Google”

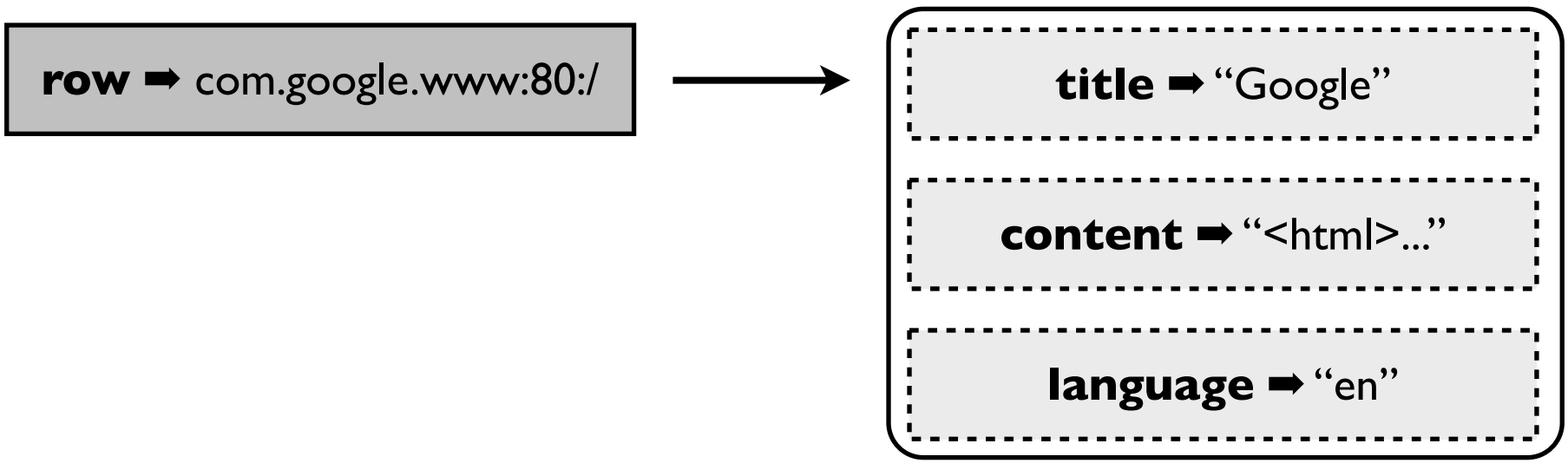
content ➡ “<html>...”

(NoSQL) Data model



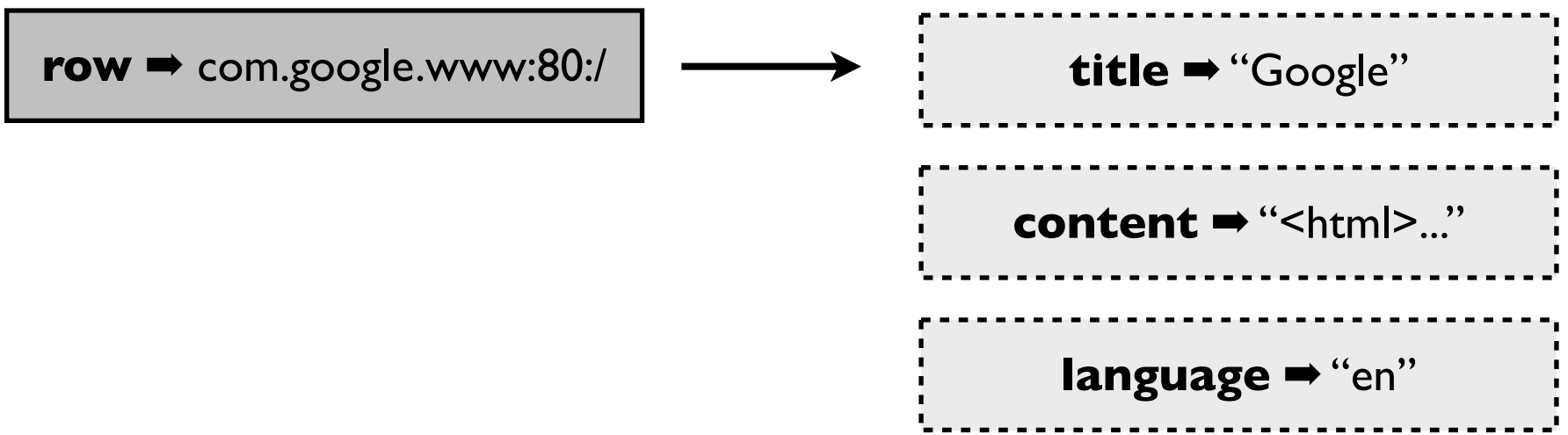
Cassandra

(NoSQL) Data model



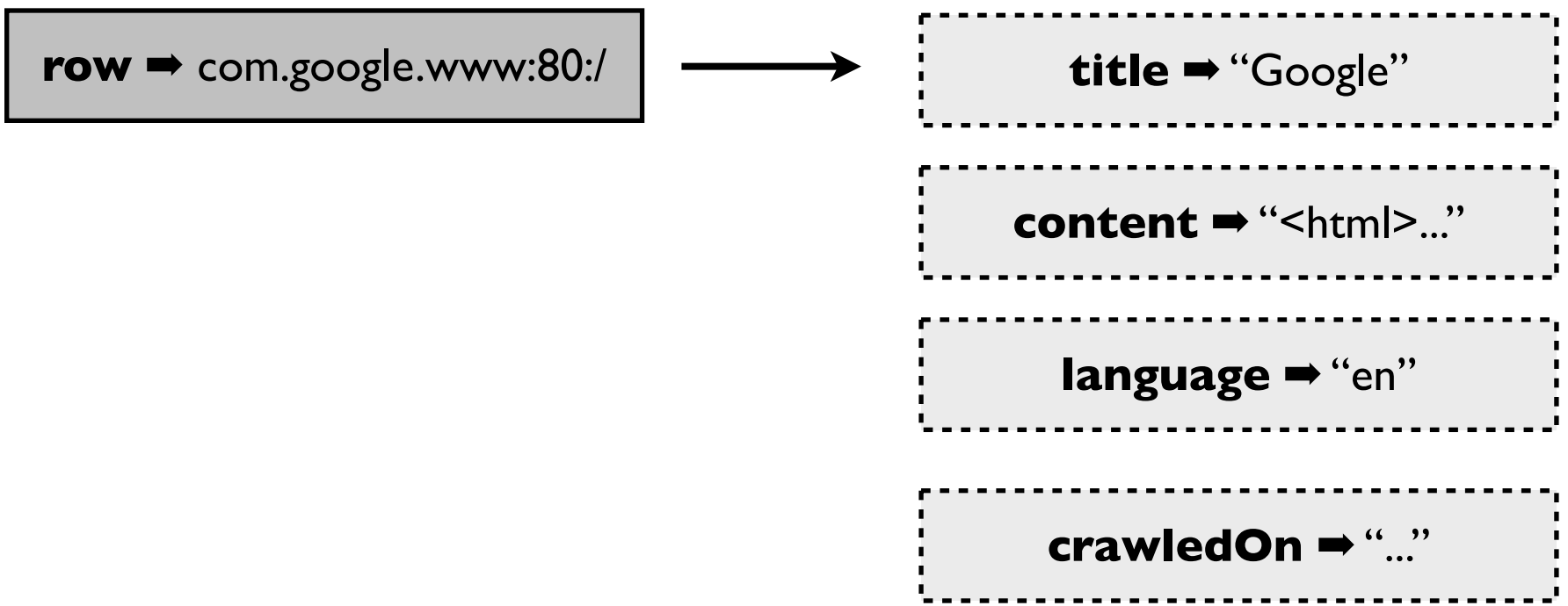
Cassandra

(NoSQL) Data model



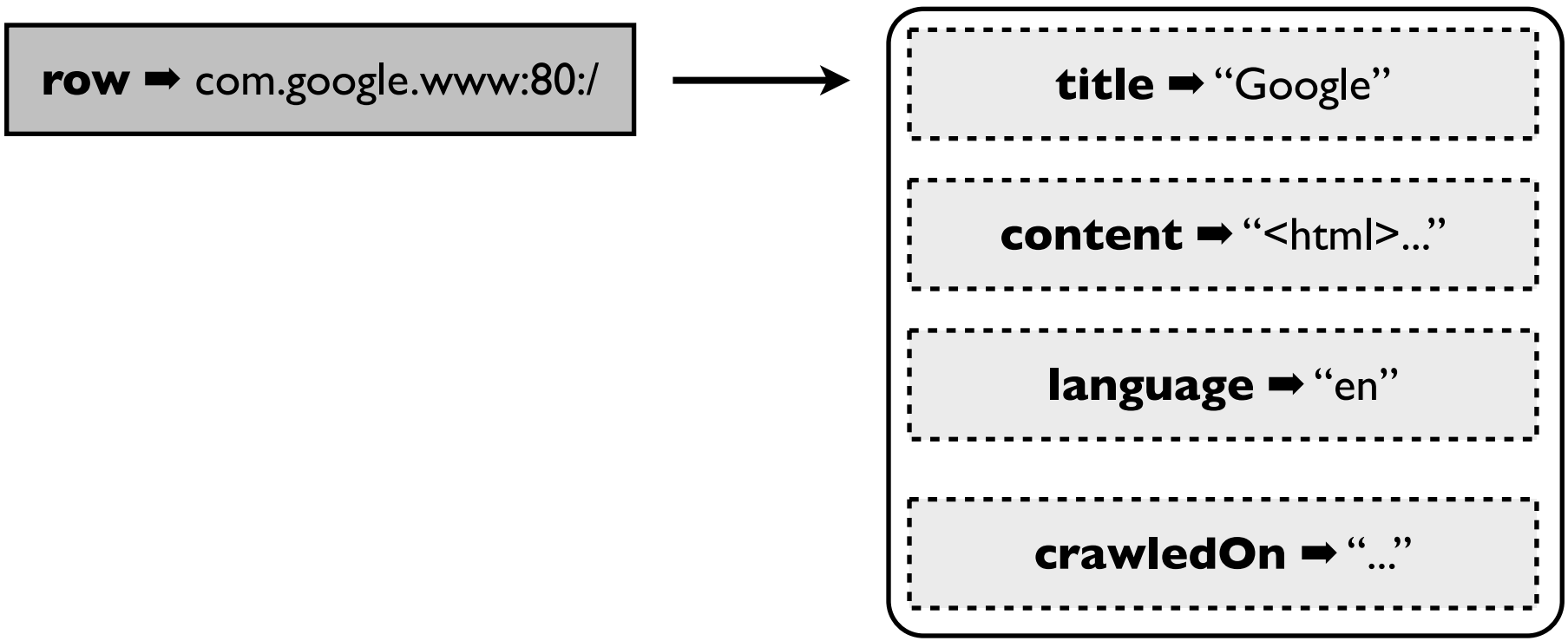
Cassandra

(NoSQL) Data model



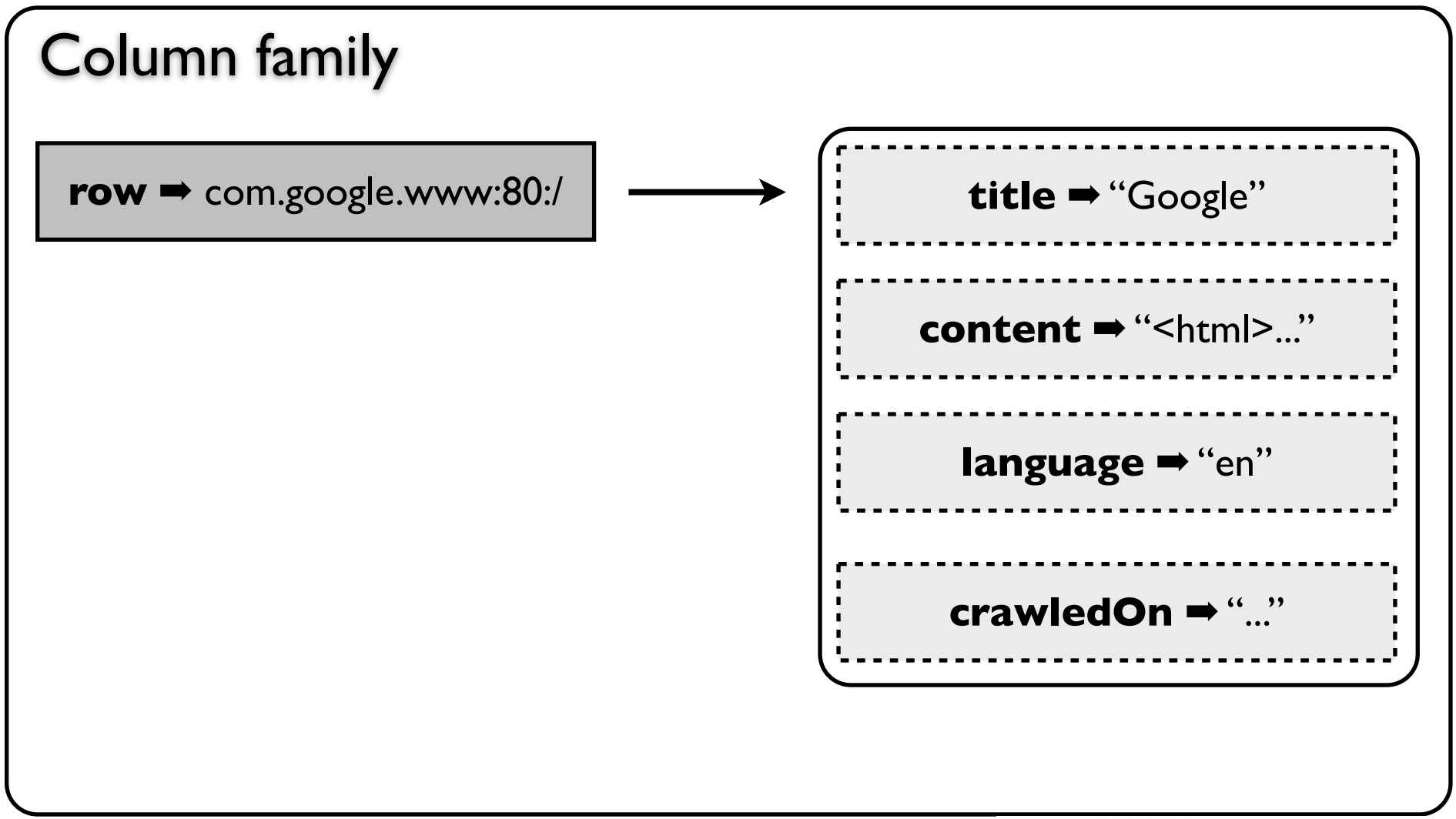
Cassandra

(NoSQL) Data model



Cassandra

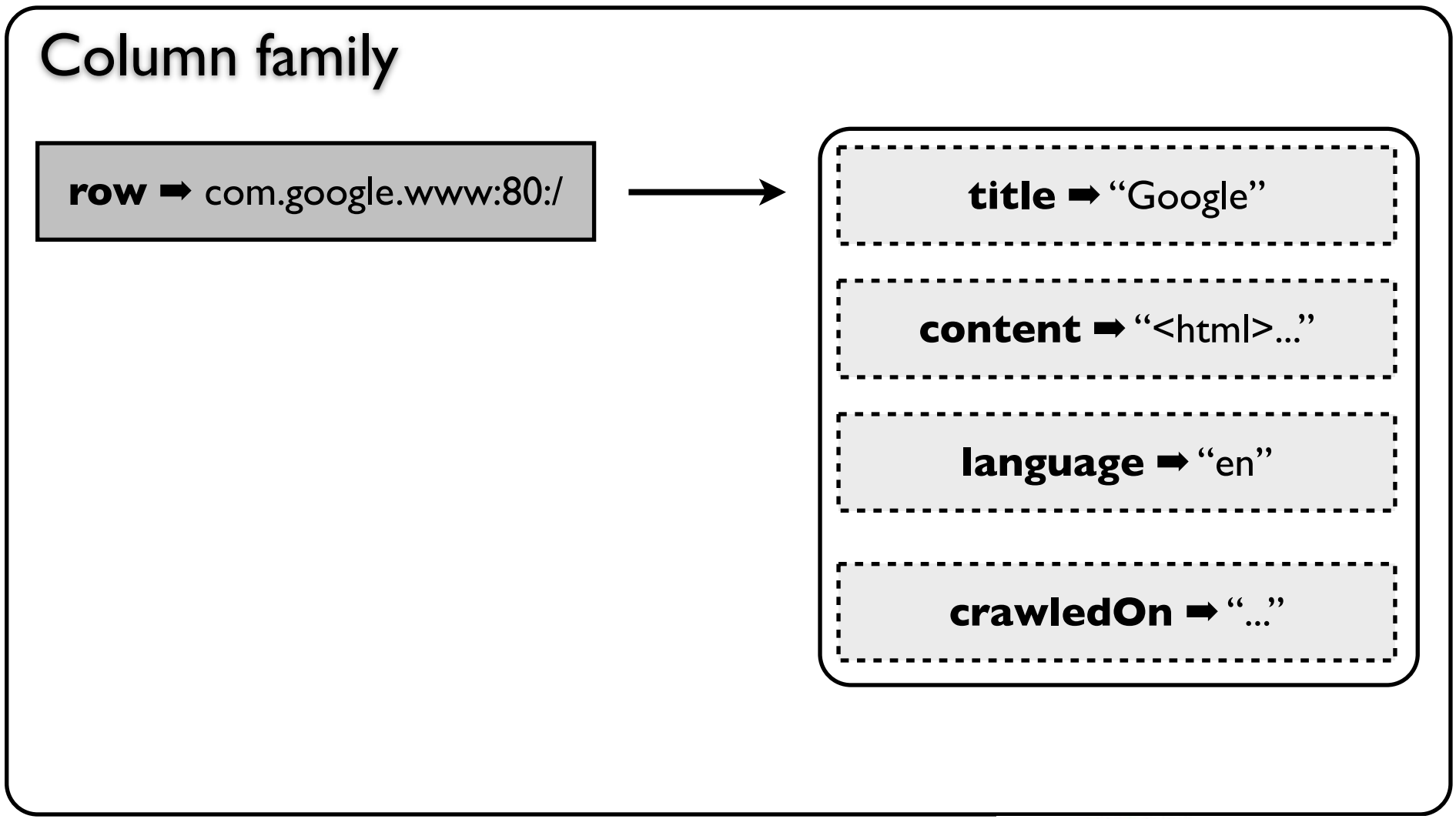
(NoSQL) Data model



Cassandra

(NoSQL) Data model

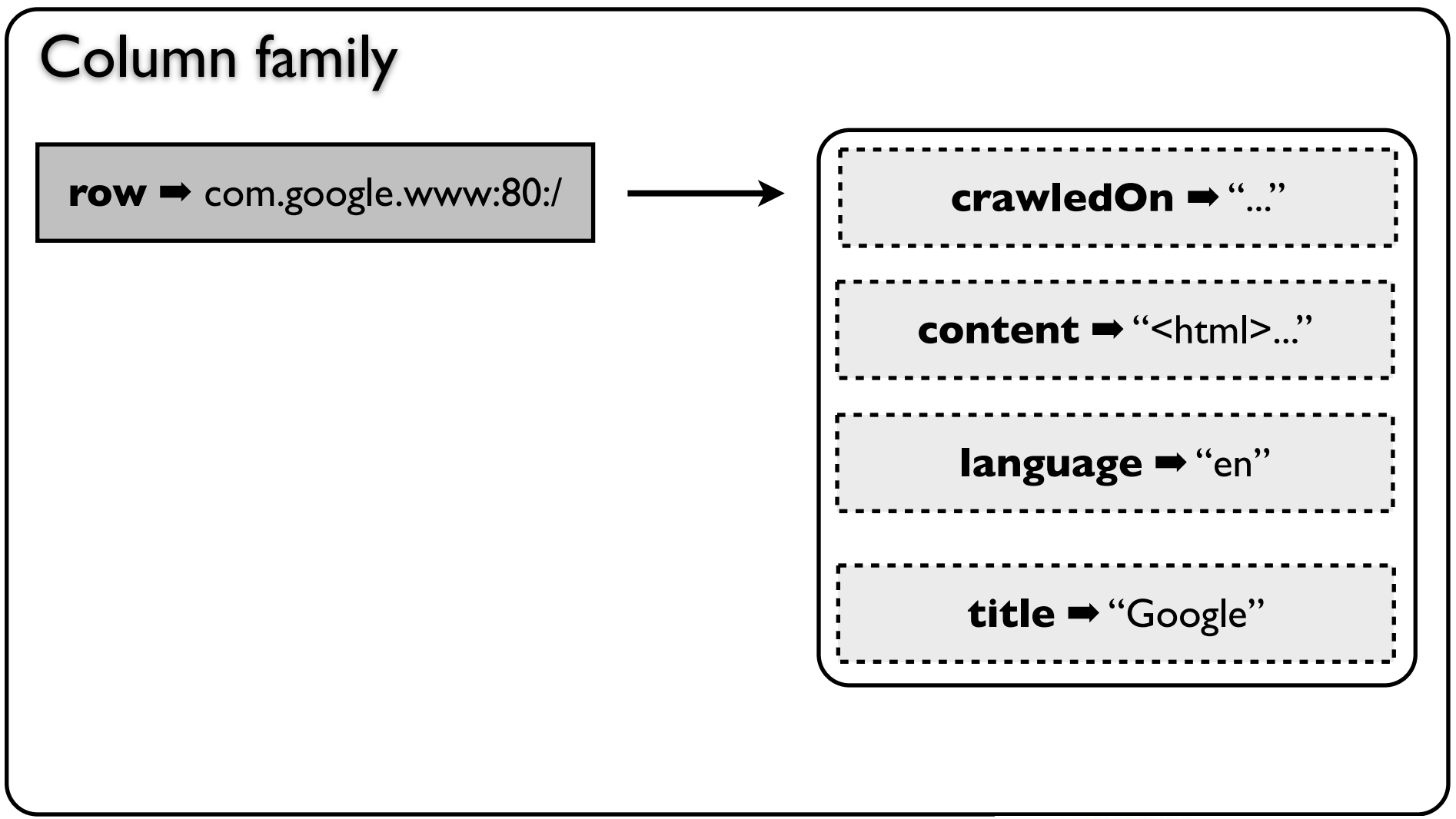
Columns are sorted



Cassandra

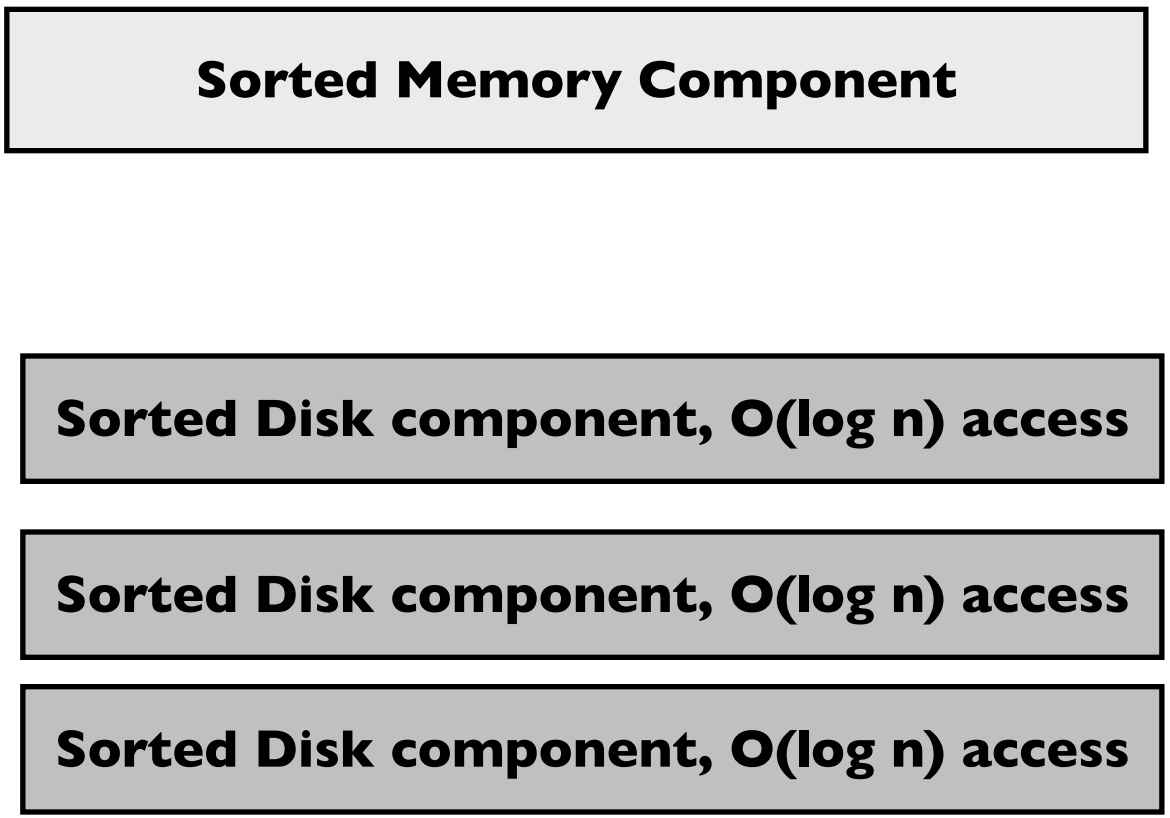
(NoSQL) Data model

Columns are sorted



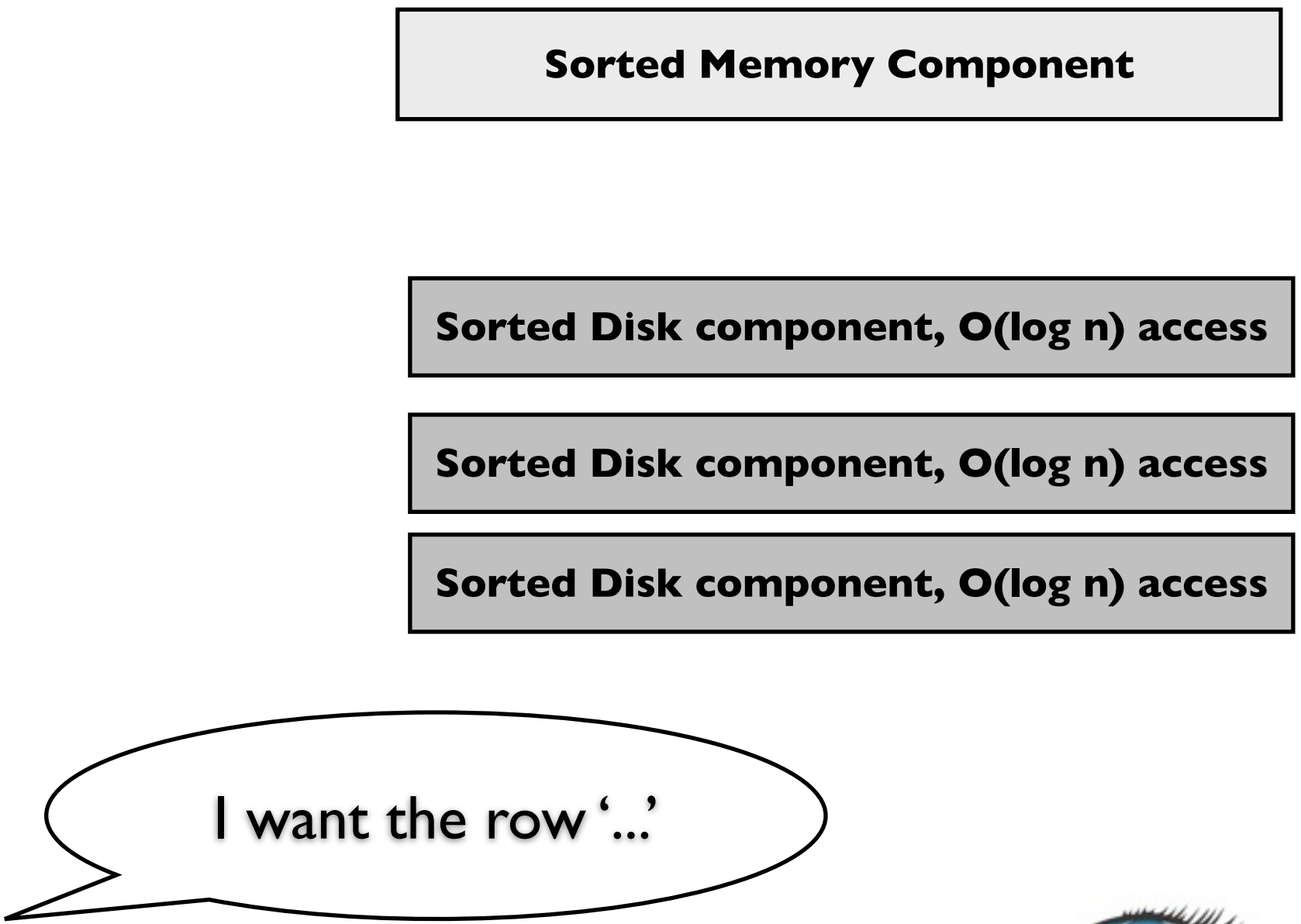
Cassandra

It is a LSM tree (Log structured merge tree)

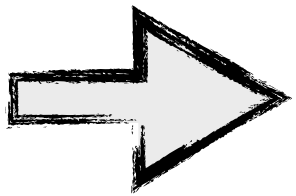


Cassandra

It is a LSM tree (Log structured merge tree)



It is a LSM tree (Log structured merge tree)



Sorted Memory Component

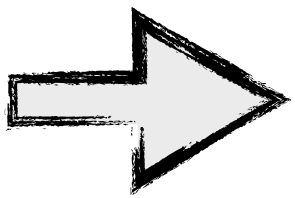
Sorted Disk component, O(log n) access

Sorted Disk component, O(log n) access

Sorted Disk component, O(log n) access

I want the row '...'

It is a LSM tree (Log structured merge tree)



Sorted Disk component, O(log n) access

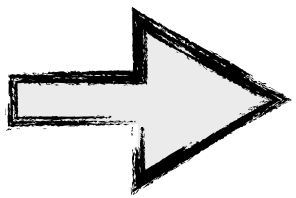
Sorted Disk component, O(log n) access

Sorted Disk component, O(log n) access

I want the row '...'

Cassandra

It is a LSM tree (Log structured merge tree)



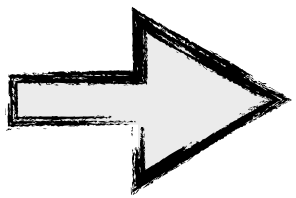
Sorted Disk component, O(log n) access

Sorted Disk component, O(log n) access

I want the row '...'

Cassandra

It is a LSM tree (Log structured merge tree)

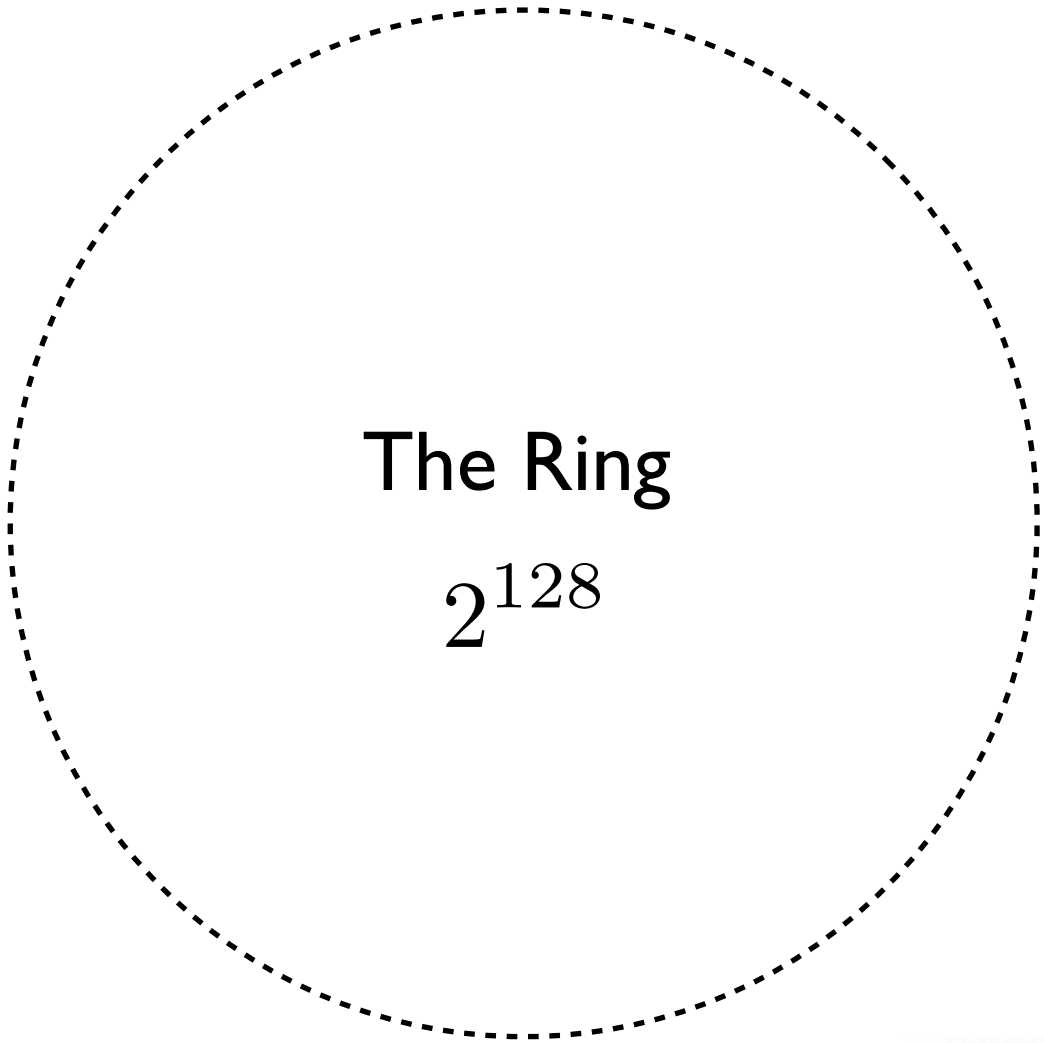


Sorted Disk component, O(log n) access

I want the row '...'

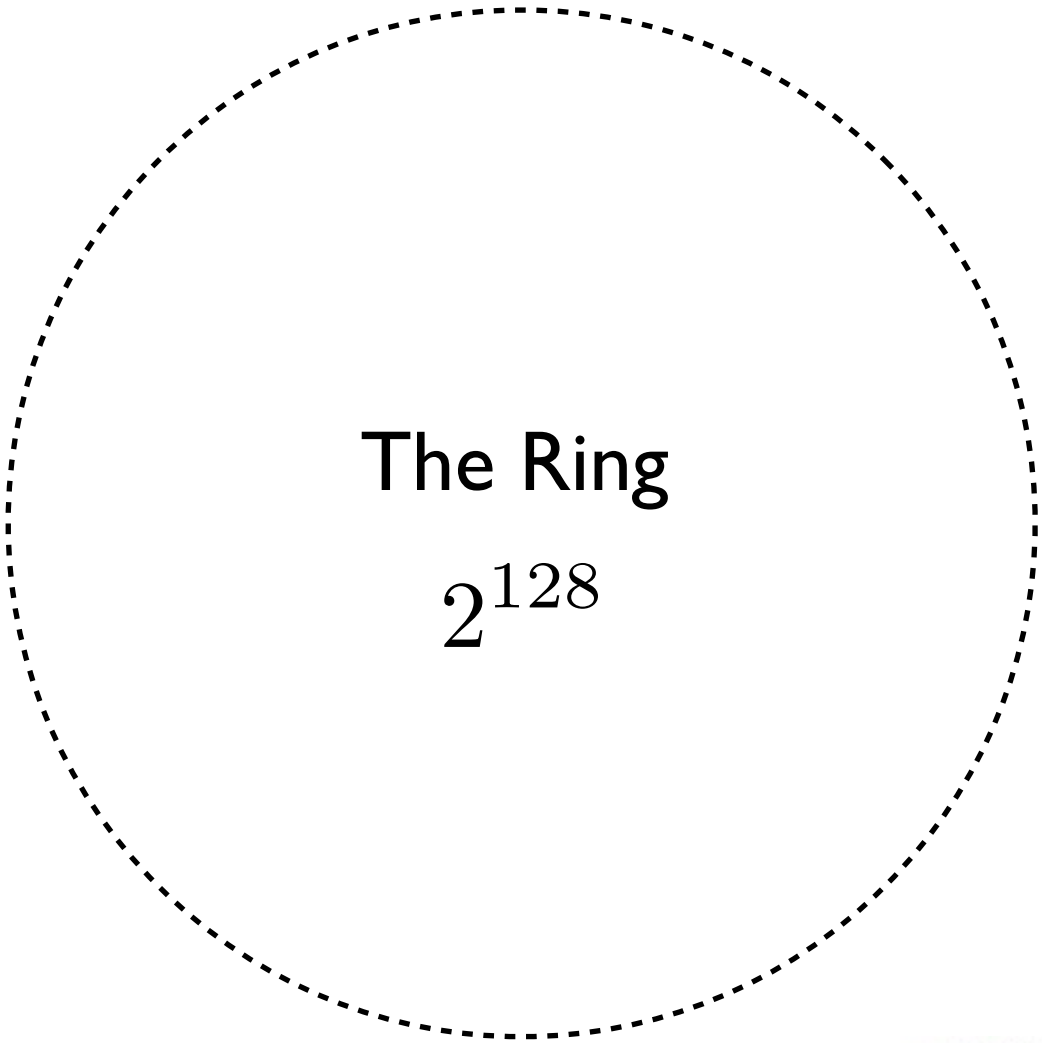
Cassandra

Data distribution



Data distribution

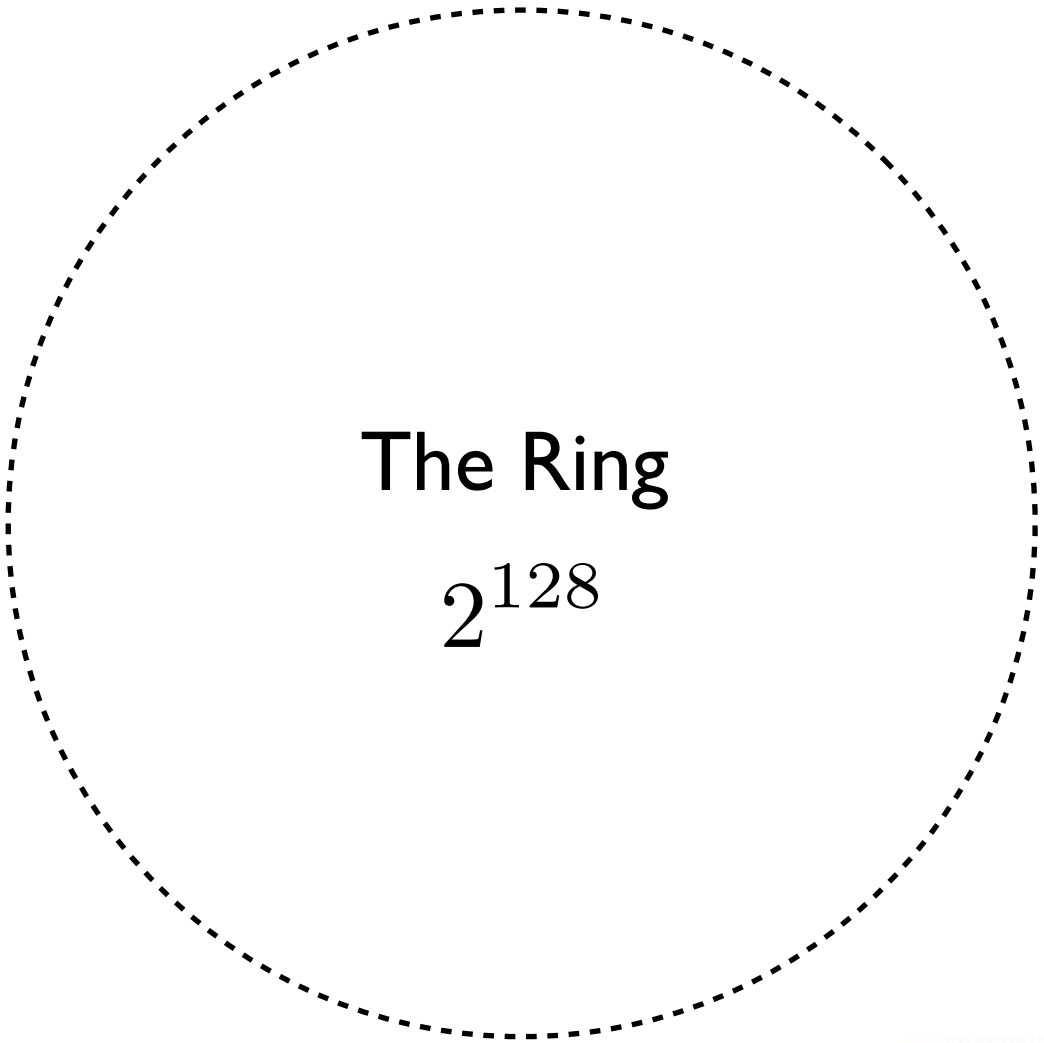
row ➡ com.google.www:80:/



Cassandra

Data distribution

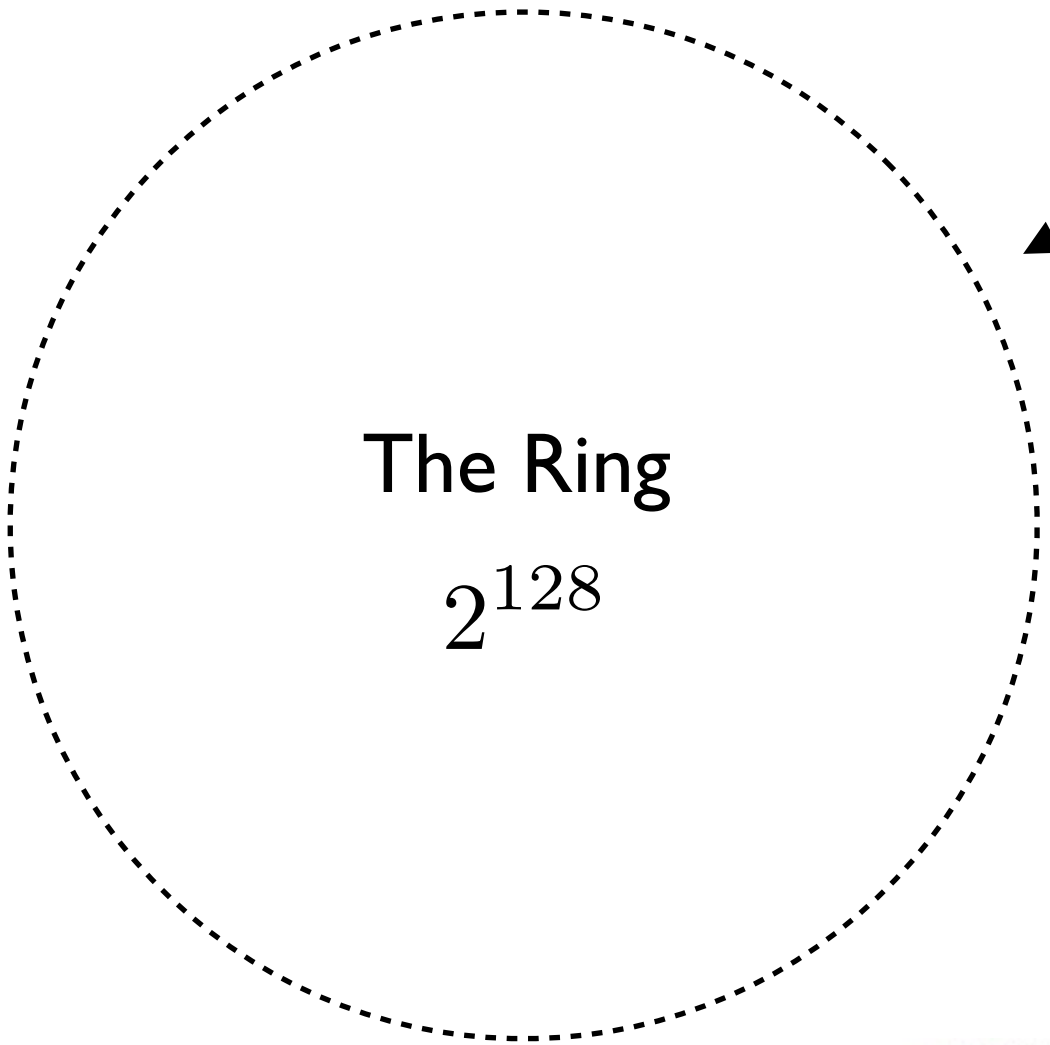
Murmur3(**row ➡** com.google.www:80:/) = ef4364eb9e...



Cassandra

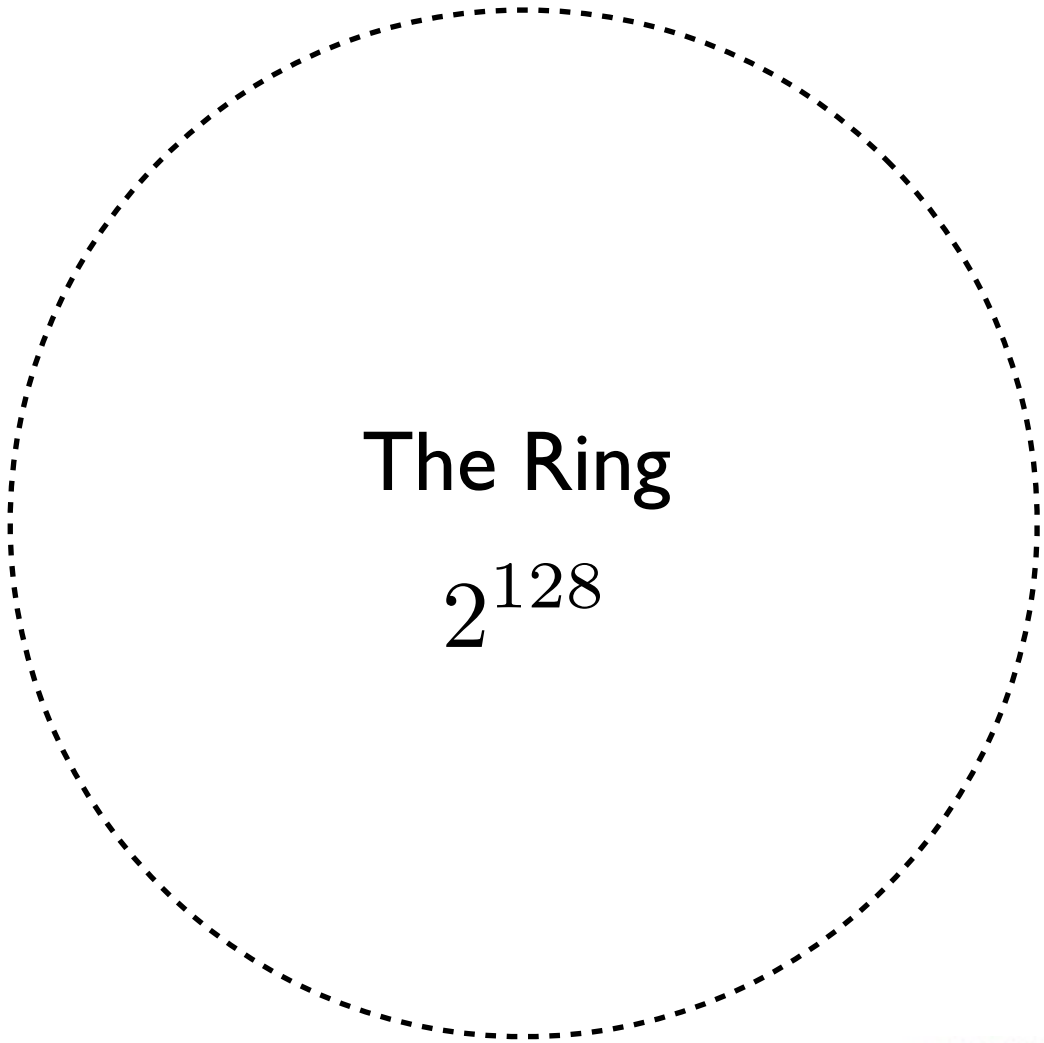
Data distribution

Murmur3(**row ➡** com.google.www:80:/) = ef4364eb9e...

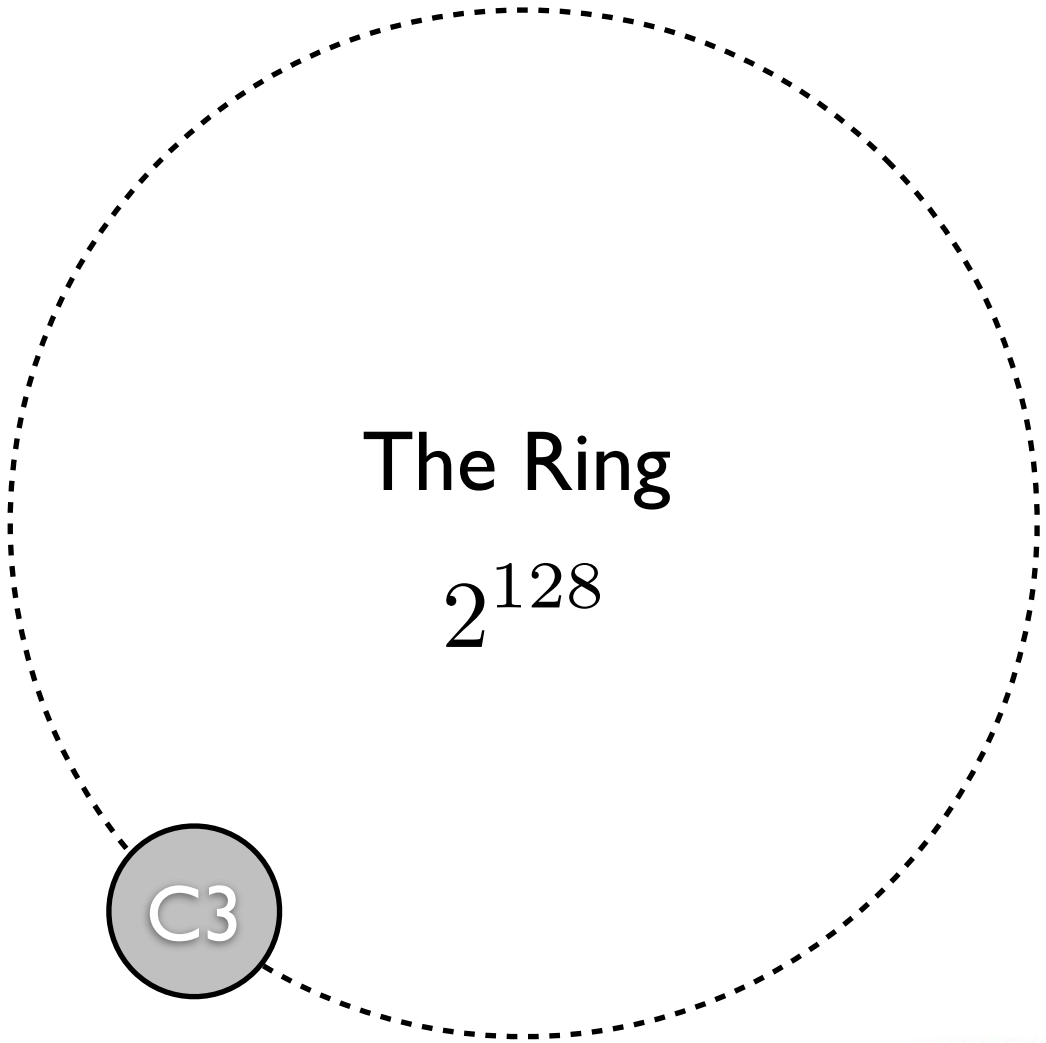


Cassandra

Data distribution

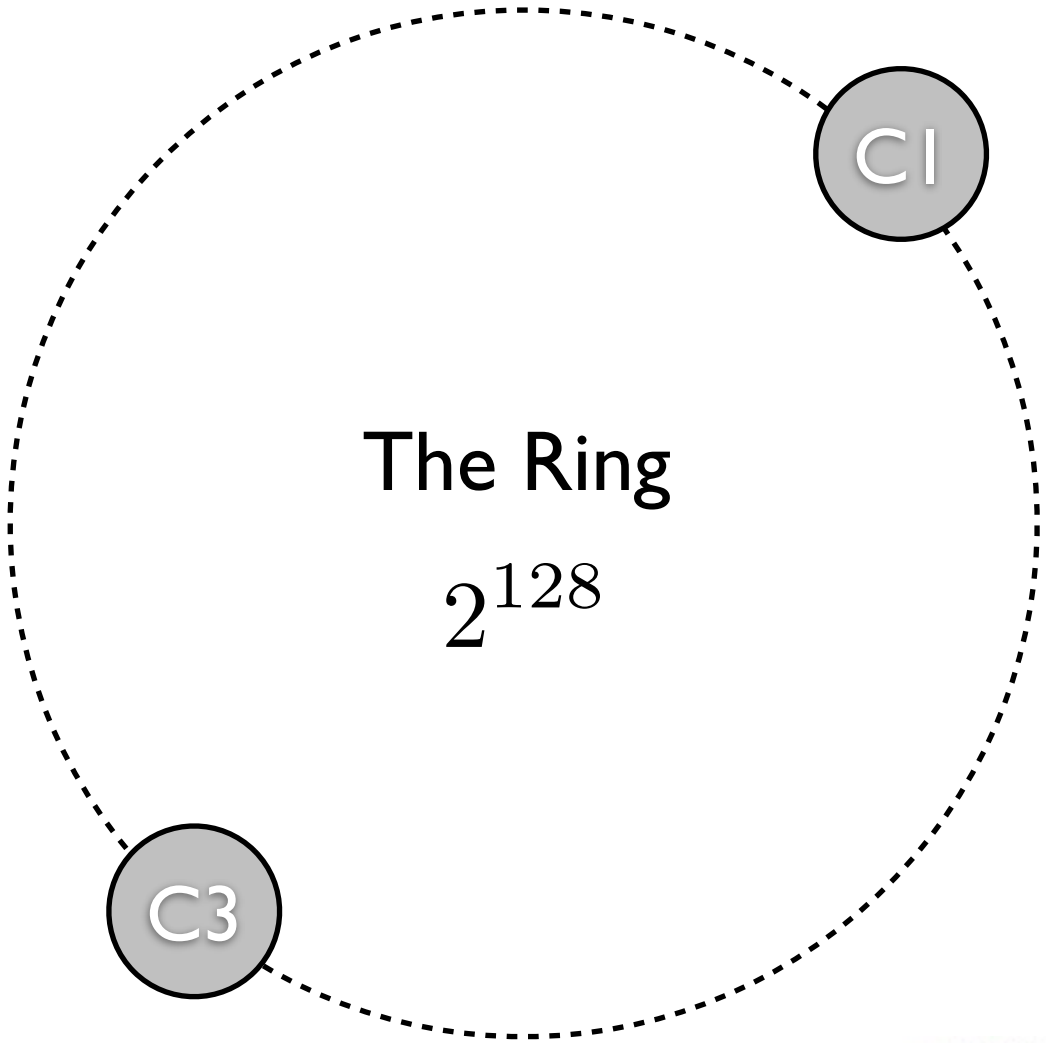


Data distribution



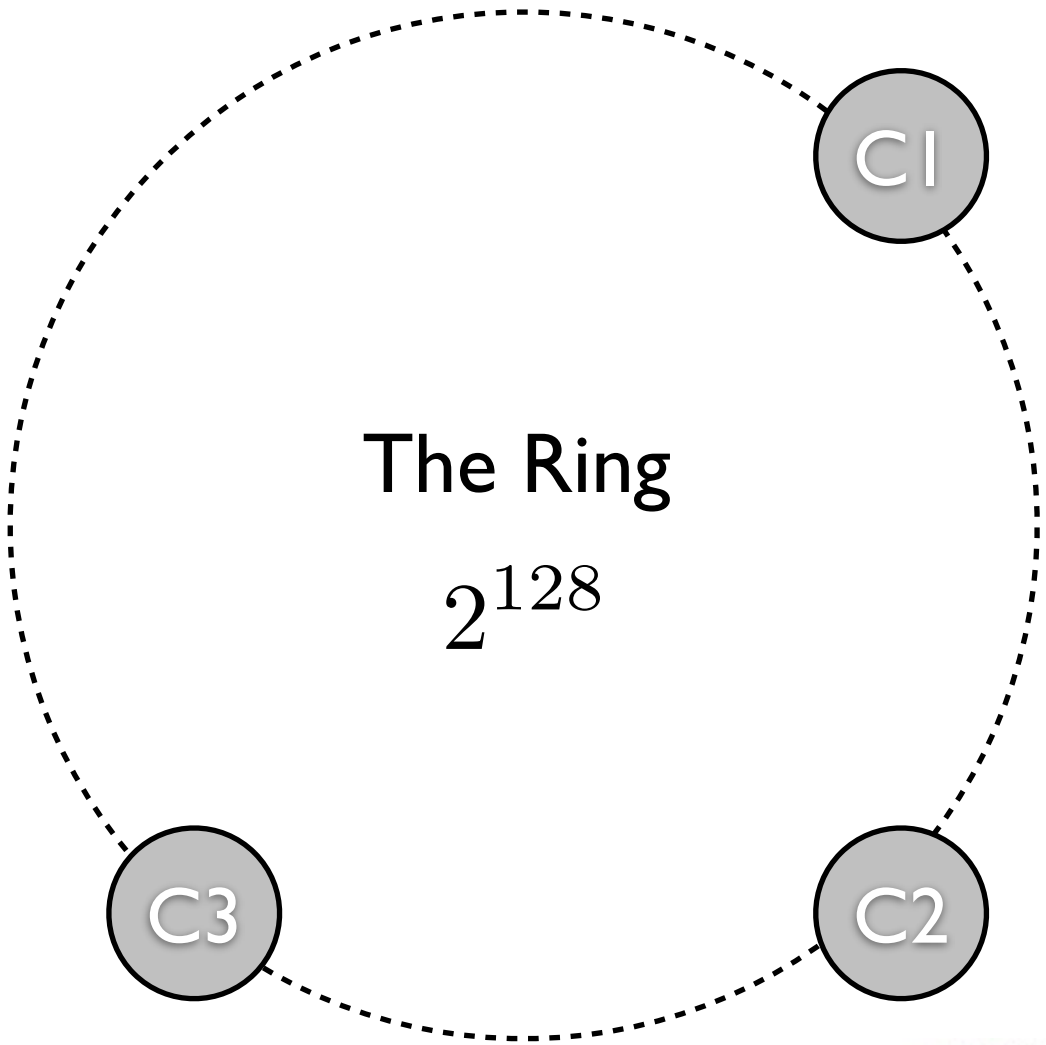
Cassandra

Data distribution



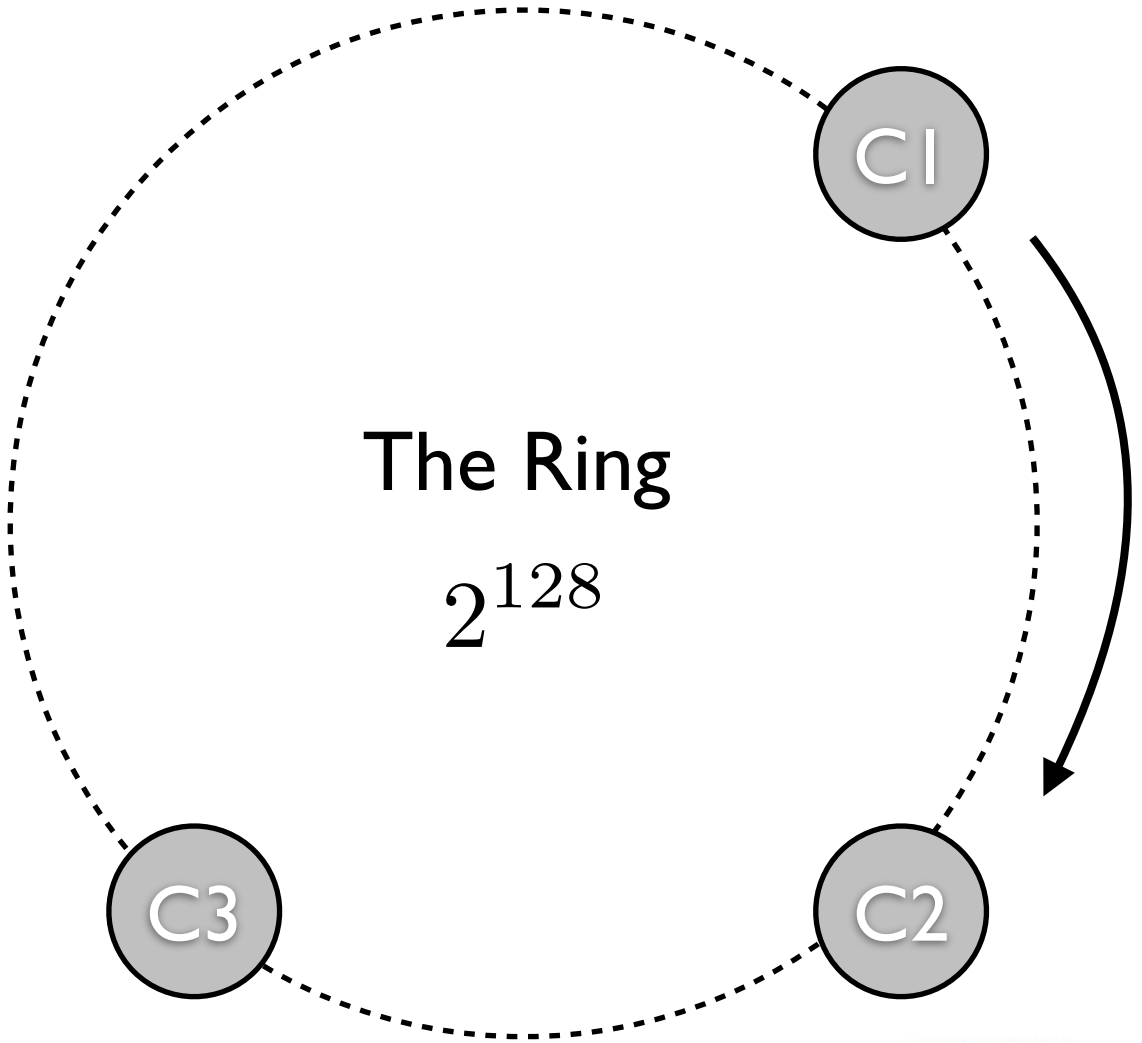
Cassandra

Data distribution



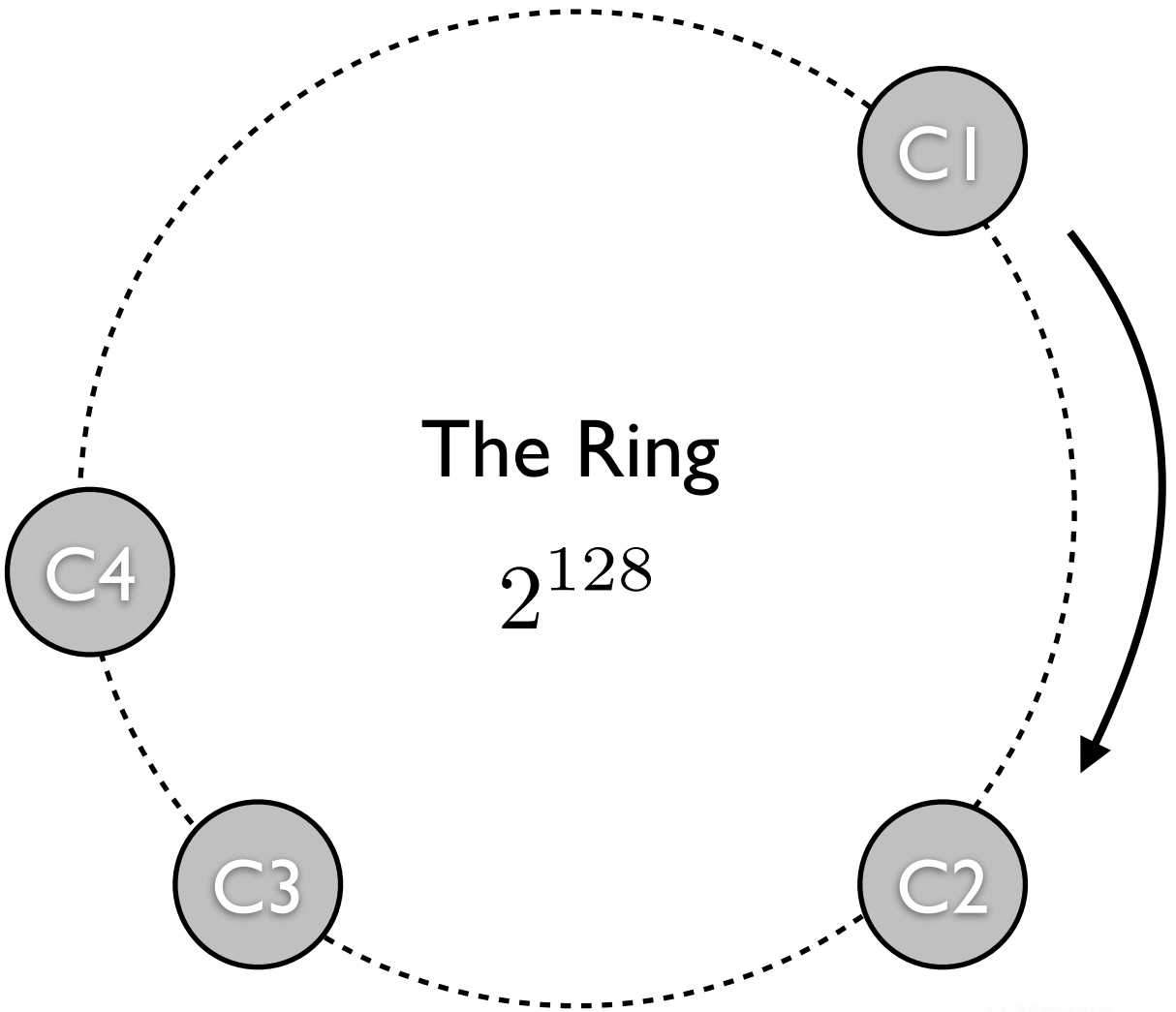
Cassandra

Data distribution



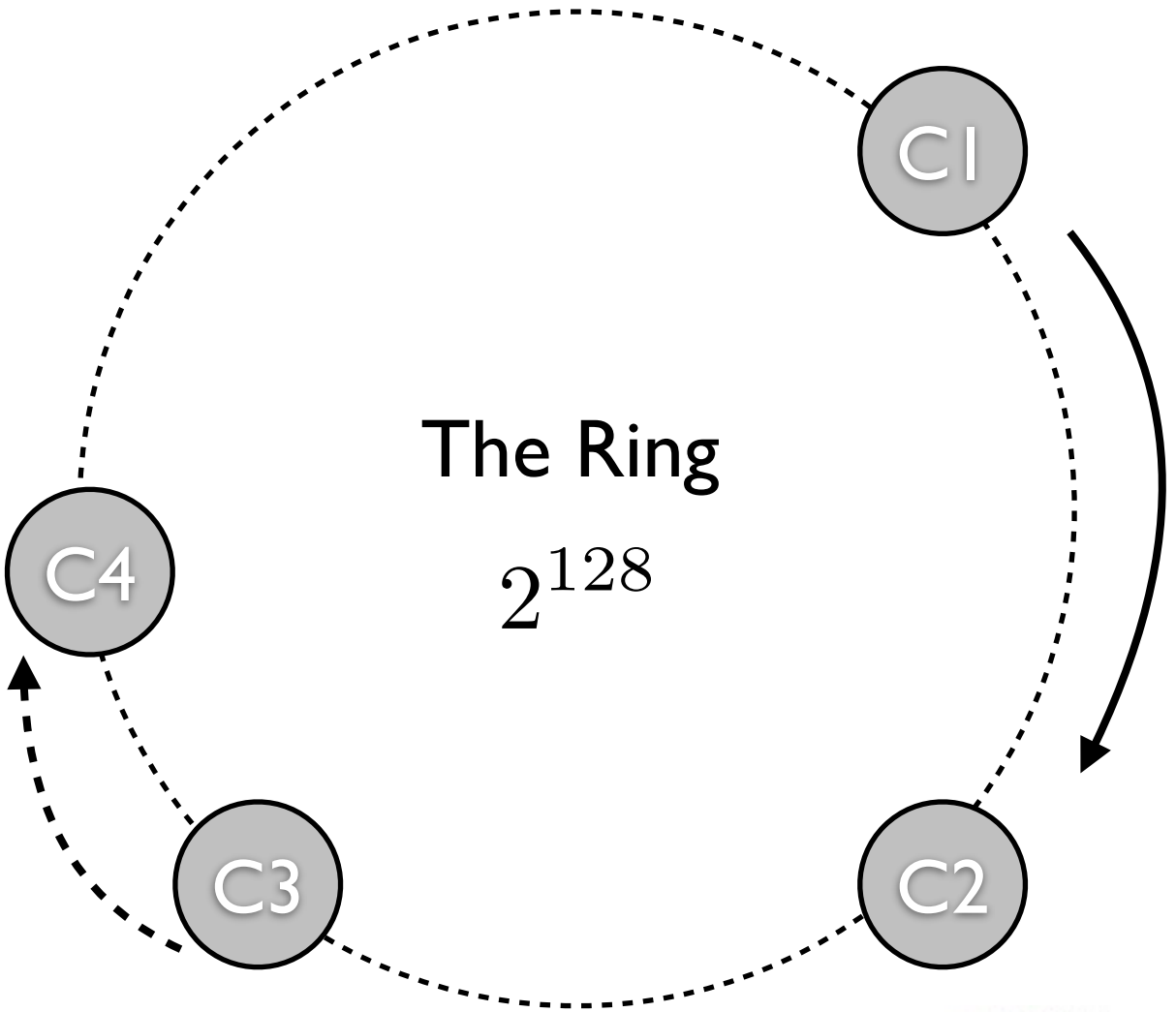
Cassandra

Data distribution



Cassandra

Data distribution



Cassandra

Demo

Post processing

Post processing

- Extract what you need from the raw HTML

Post processing

- Extract what you need from the raw HTML
  - Encoding (Unicode hell)

Post processing

- Extract what you need from the raw HTML
  - Encoding (Unicode hell)
  - Language

Post processing

- Extract what you need from the raw HTML
  - Encoding (Unicode hell)
  - Language
  - Date

Post processing

- Extract what you need from the raw HTML
  - Encoding (Unicode hell)
  - Language
  - Date
  - Title

Post processing

- Extract what you need from the raw HTML
  - Encoding (Unicode hell)
  - Language
  - Date
  - Title
  - Content of the page (possibly cleaned up)

Post processing

- Extract what you need from the raw HTML
  - Encoding (Unicode hell)
  - Language
  - Date
  - Title
  - Content of the page (possibly cleaned up)
  - Description, keywords

Post processing

- Extract what you need from the raw HTML
  - Encoding (Unicode hell)
  - Language
  - Date
  - Title
  - Content of the page (possibly cleaned up)
  - Description, keywords
  - Out links

Post processing

- Extract what you need from the raw HTML
  - Encoding (Unicode hell)
  - Language
  - Date
  - Title
  - Content of the page (possibly cleaned up)
  - Description, keywords
  - Out links
  - Images, videos, favicons?

Post processing

- Extract what you need from the raw HTML
  - Encoding (Unicode hell)
  - Language
  - Date
  - Title
  - Content of the page (possibly cleaned up)
  - Description, keywords
  - Out links
  - Images, videos, favicons?
  - NLP and topic annotator tools (à la TagMe, here applied on title)

Post processing

- Extract what you need from the raw HTML
  - Encoding (Unicode hell)
  - Language
  - Date
  - Title
  - Content of the page (possibly cleaned up)
  - Description, keywords
  - Out links
  - Images, videos, favicons?
  - NLP and topic annotator tools (à la TagMe, here applied on title)

Post processing

- Extract what you need from the raw HTML
  - Encoding (Unicode hell)
  - Language
  - Date
  - Title
  - Content of the page (possibly cleaned up)
  - Description, keywords
  - Out links

  First gen
  - Images, videos, favicons?
  - NLP and topic annotator tools (à la TagMe, here applied on title)

Post processing

- Extract what you need from the raw HTML
  - Encoding (Unicode hell)
  - Language
  - Date
  - Title
  - Content of the page (possibly cleaned up)
  - Description, keywords
  - Out links

First gen

PageRank

  - Images, videos, favicons?
  - NLP and topic annotator tools (à la TagMe, here applied on title)

Link analysis with Spark

Link analysis with Spark

- Why?

Link analysis with Spark

- Why?
  - In-memory computing

Link analysis with Spark

- Why?
  - In-memory computing
  - Distributed

Link analysis with Spark

- Why?
  - In-memory computing
  - Distributed
  - Fast to write and to run

Link analysis with Spark

- Why?
  - In-memory computing
  - Distributed
  - Fast to write and to run
  - You can use Hadoop (DFS), etc..

Link analysis with Spark

- Why?
  - In-memory computing
  - Distributed
  - Fast to write and to run
  - You can use Hadoop (DFS), etc..
  - Open source

Link analysis with Spark

- Why?
  - In-memory computing
  - Distributed
  - Fast to write and to run
  - You can use Hadoop (DFS), etc..
  - Open source
  - Python, Scala, Java

Link analysis with Spark

- Why?
  - In-memory computing
  - Distributed
  - Fast to write and to run
  - You can use Hadoop (DFS), etc..
  - Open source
  - Python, Scala, Java
  - Easy model to work with

Link analysis with Spark

- Why?
  - In-memory computing
  - Distributed
  - Fast to write and to run
  - You can use Hadoop (DFS), etc..
  - Open source
  - Python, Scala, Java
  - Easy model to work with

Suitable for iterative algorithms running in memory
(Machine Learning, PageRank, ...)

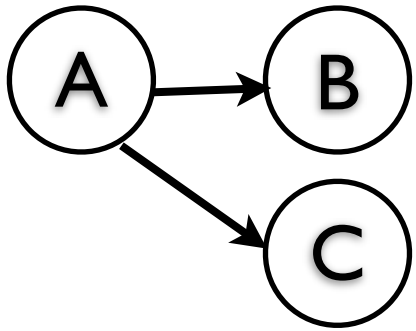
Spark building block

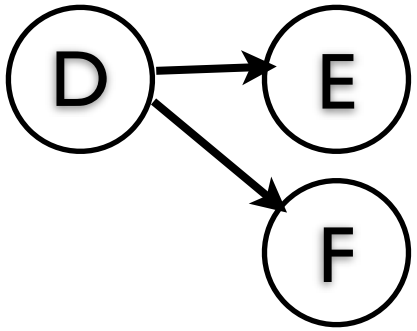
- Spark uses RDDs (Resilient Distributed Datasets)
  - A read only partitioned collection of records
  - Transform RDDs in RDDs (not in-place, aka functional)
  - Useful operations:
    - `map(f: T => U) : RDD[T] => RDD[U]`
    - `groupByKey() : RDD[(K,V)] => RDD[(K, Seq[V])]`
    - `mapValues(f: V => W) : RDD[(K,V)] => RDD[(K,W)]`
    - `join() : RDD[(K,V)], RDD[(K,W)] => RDD[(K, (V,W))]`
    - `reduceByKey(f: (V,V) => V) : RDD[(K,V)] => RDD[(K,V)]`

Simil-*PageRank* in Spark

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

http://example.org	http://whatever.org
http://example.org	http://donthaveasite.org
http://repubblica.it	http://repubblica.it/video
http://repubblica.it	http://it.wikipedia.org/wiki/..

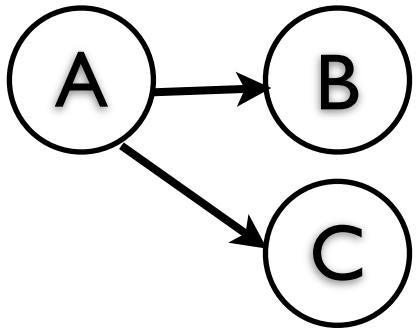


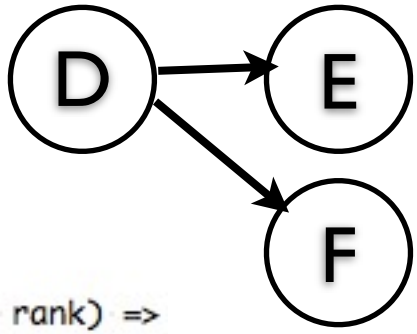


Simil-*PageRank* in Spark

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

http://example.org http://whatever.org
http://example.org http://donthaveasite.org
http://repubblica.it http://repubblica.it/video
http://repubblica.it http://it.wikipedia.org/wiki/..



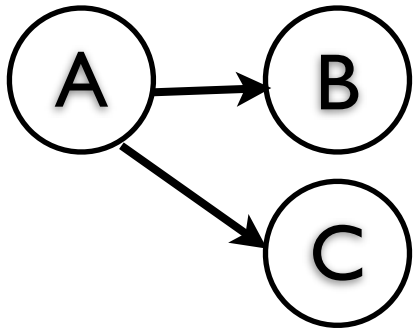


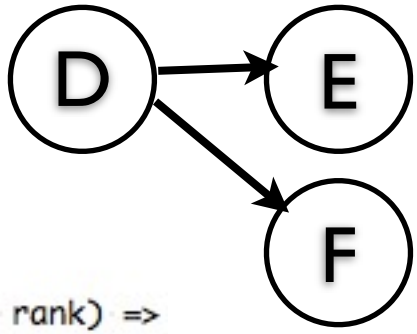

```
val lines = ctx.textFile(args(1), 1)
val links = lines.map{ s =>
  val parts = s.split("\\s+")
  (parts(0), parts(1))
}.distinct().groupByKey().cache()
var ranks = links.mapValues(v => 1.0)

for (i <- 1 to iters) {
  val contribs = links.join(ranks).values.flatMap{ case (urls, rank) =>
    val size = urls.size
    urls.map(url => (url, rank / size))
  }
  ranks = contribs.reduceByKey(_ + _).mapValues(0.15 + 0.85 * _)
}

val output = ranks.collect()
output.foreach(tup => println(tup._1 + " has rank: " + tup._2 + "."))
```

Simil-*PageRank* in Spark

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

http://example.org http://whatever.org
http://example.org http://donthaveasite.org
http://repubblica.it http://repubblica.it/video
http://repubblica.it http://it.wikipedia.org/wiki/..

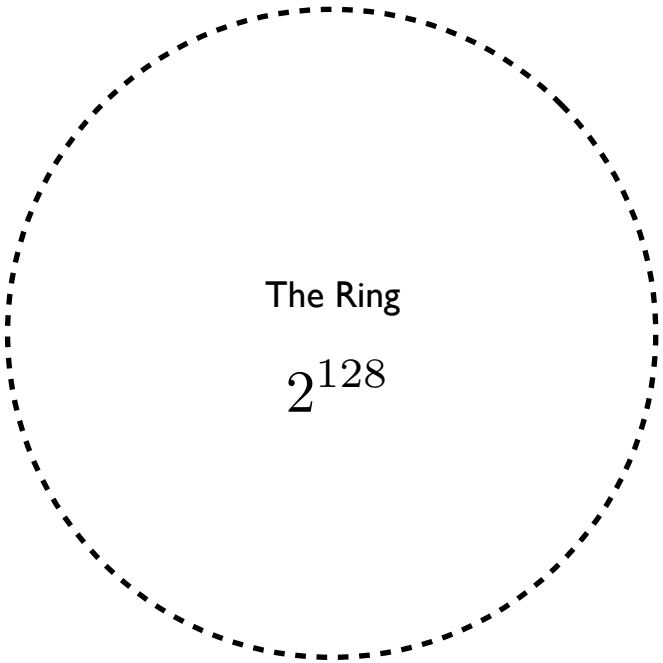




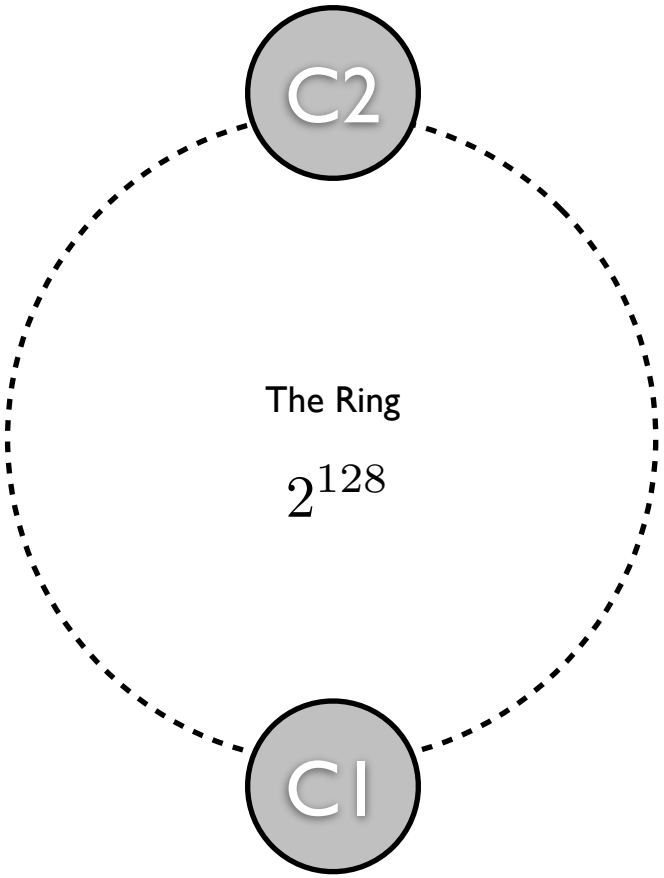

```
val lines = ctx.textFile(args(1), 1)
val links = lines.map{ s =>
  val parts = s.split("\\s+")
  (parts(0), parts(1))
}.distinct().groupByKey().cache()
var ranks = links.mapValues(v => 1.0)

for (i <- 1 to iters) {
  val contribs = links.join(ranks).values.flatMap{ case (urls, rank) =>
    val size = urls.size
    urls.map(url => (url, rank / size))
  }
  ranks = contribs.reduceByKey(_ + _).mapValues(0.15 + 0.85 * _)
}

val output = ranks.collect()
output.foreach(tup => println(tup._1 + " has rank: " + tup._2 + "."))
```

$\frac{}{N}$

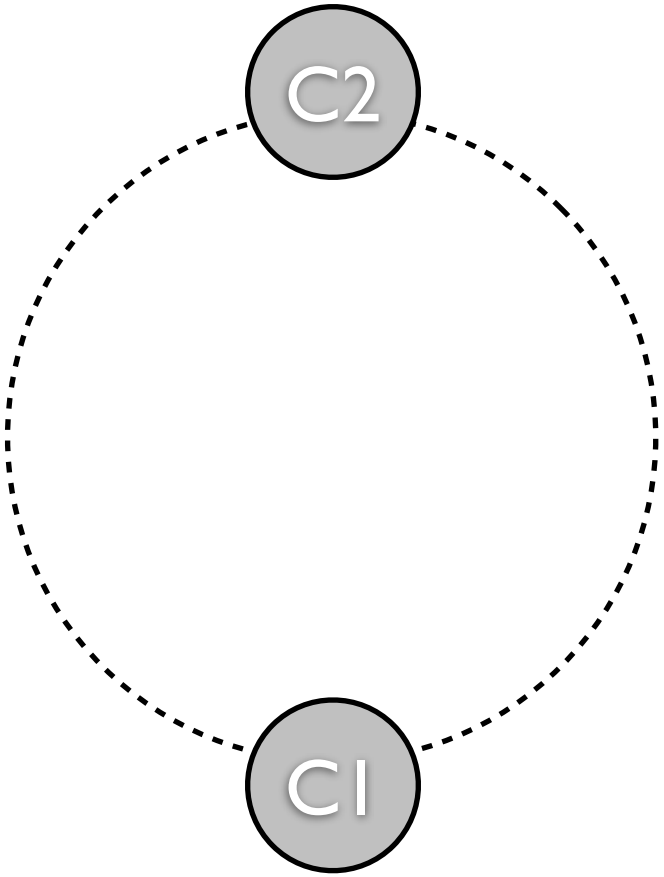
Get the graph out of Cassandra



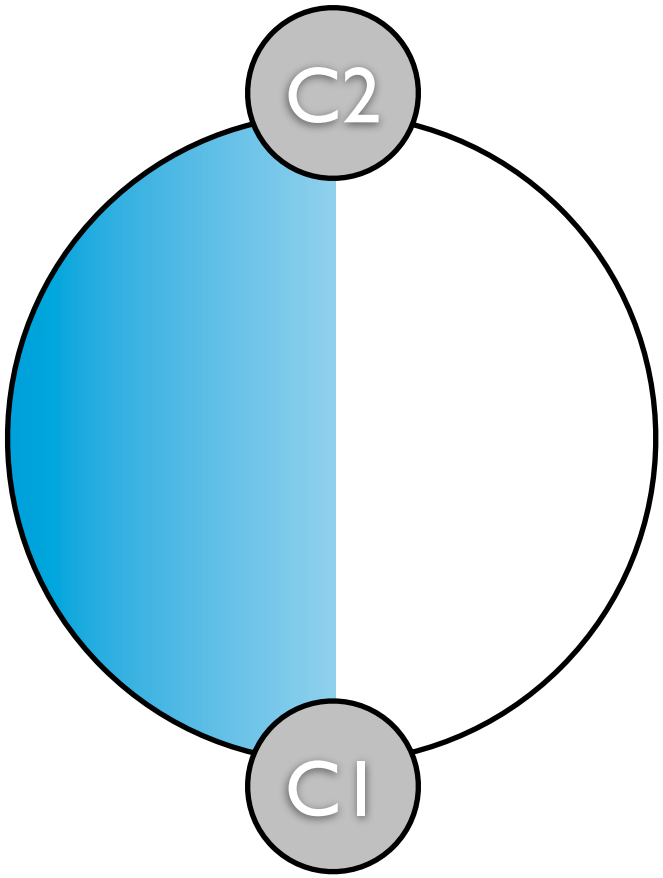
Get the graph out of Cassandra



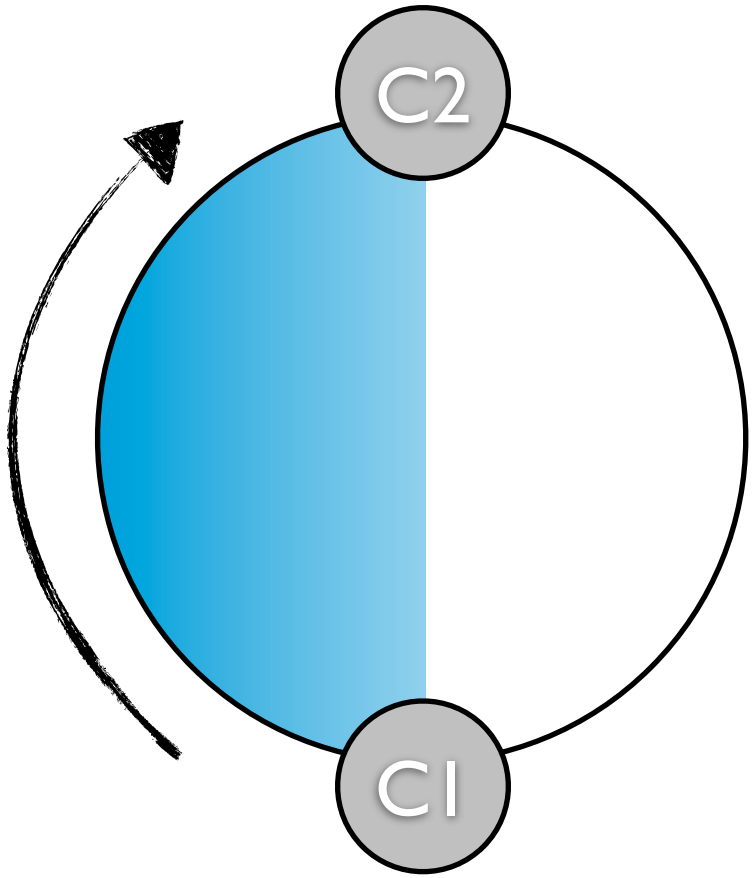
Get the graph out of Cassandra



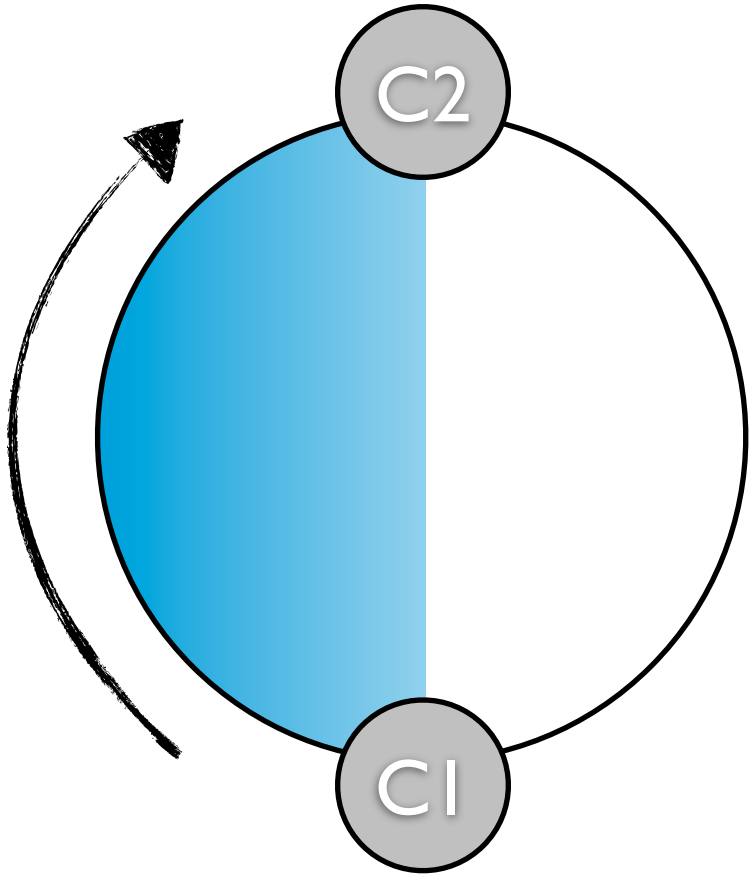
Get the graph out of Cassandra



Get the graph out of Cassandra

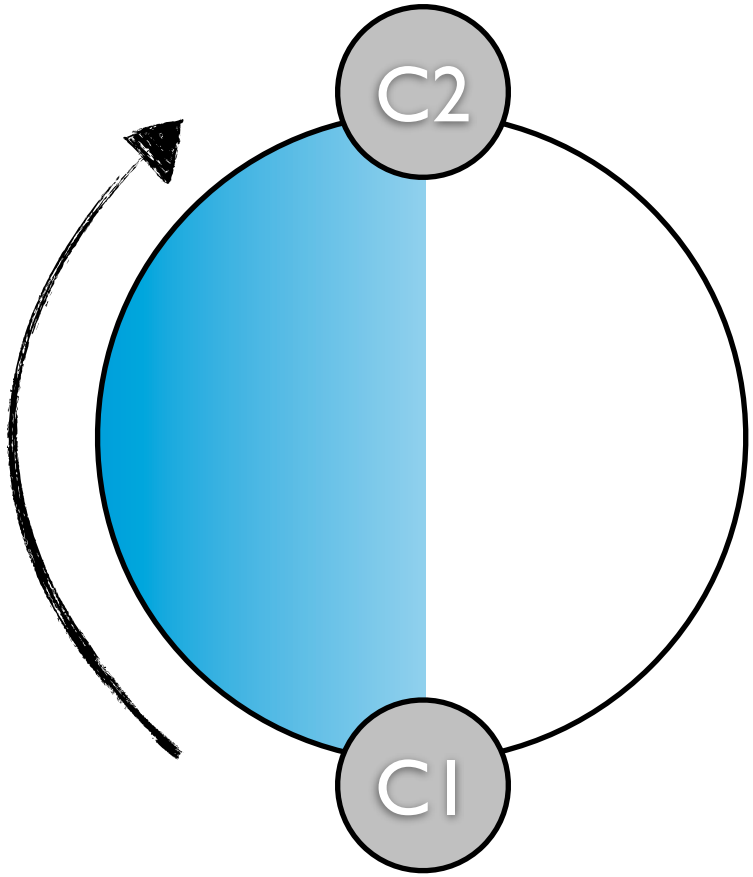


Get the graph out of Cassandra

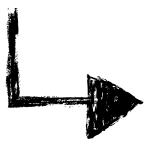


https://27esimaora.corriere.it/ ➡ https://27esimaora.corriere.it/articolo/bertolucci-e-la-scena-del..
https://27esimaora.corriere.it/ ➡ https://27esimaora.corriere.it/articolo/divorzio-allitaliana-ecco-per..
https://27esimaora.corriere.it/ ➡ https://27esimaora.corriere.it/articolo/donne-ingegnere-cinque-volte-meno-pag..
https://27esimaora.corriere.it/ ➡ https://27esimaora.corriere.it/articolo/donne-liberal-alla-conquista-del-sud..

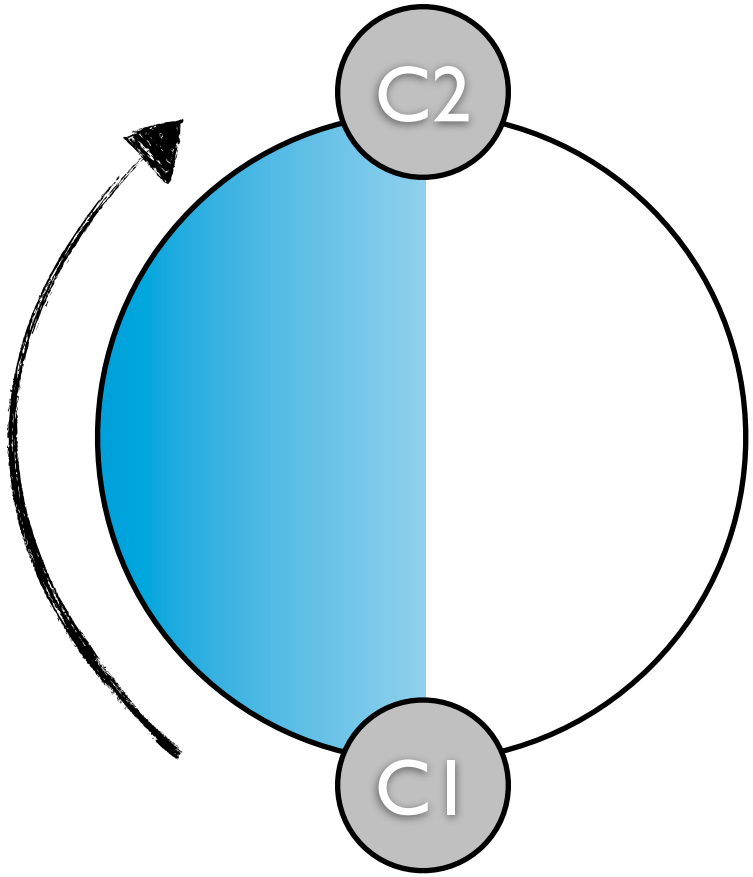
Get the graph out of Cassandra



https://27esimaora.corriere.it/ ➡ https://27esimaora.corriere.it/articolo/bertolucci-e-la-scena-del..
https://27esimaora.corriere.it/ ➡ https://27esimaora.corriere.it/articolo/divorzio-allitaliana-ecco-per..
https://27esimaora.corriere.it/ ➡ https://27esimaora.corriere.it/articolo/donne-ingegnere-cinque-volte-meno-pag..
https://27esimaora.corriere.it/ ➡ https://27esimaora.corriere.it/articolo/donne-liberal-alla-conquista-del-sud..



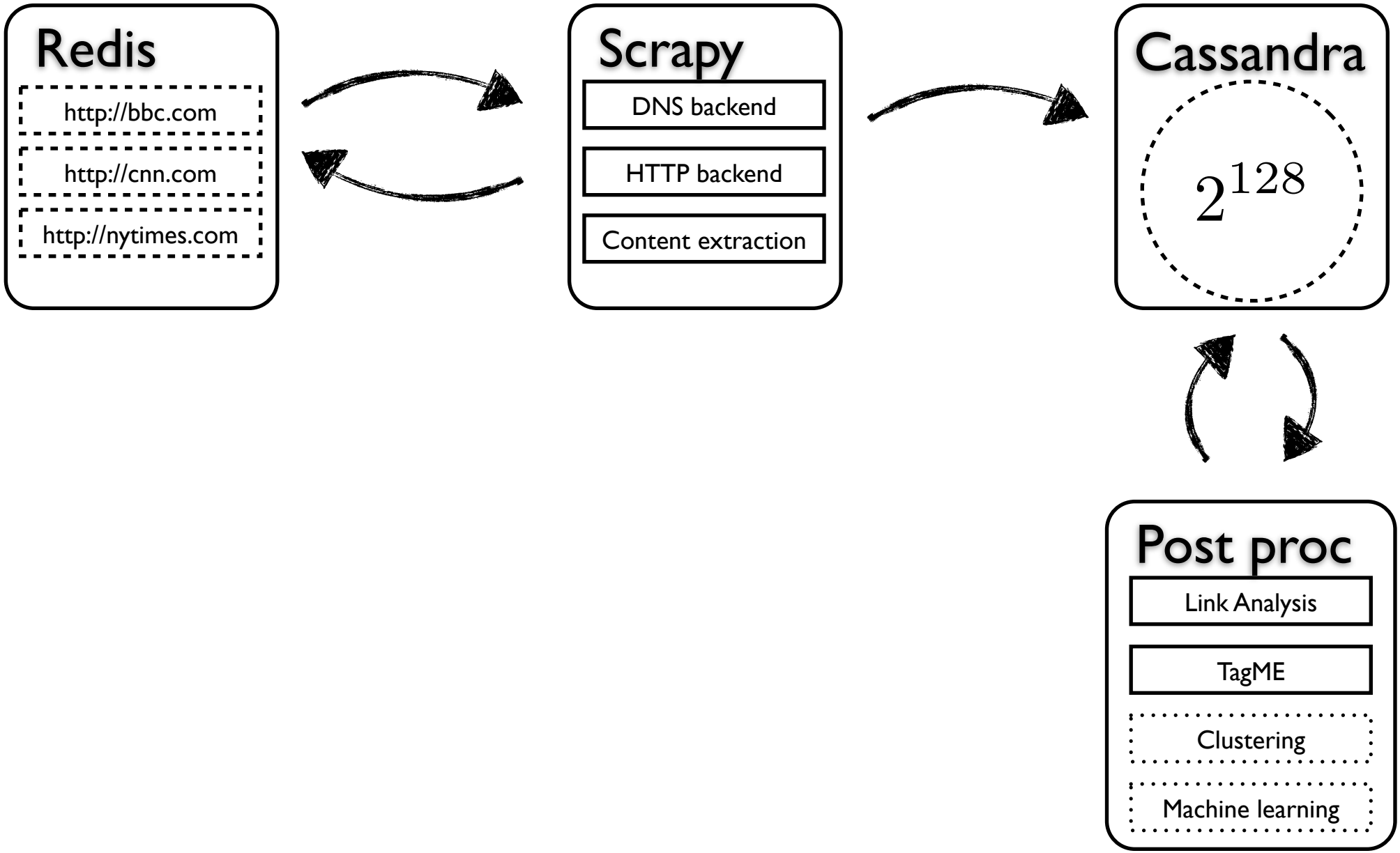
Get the graph out of Cassandra



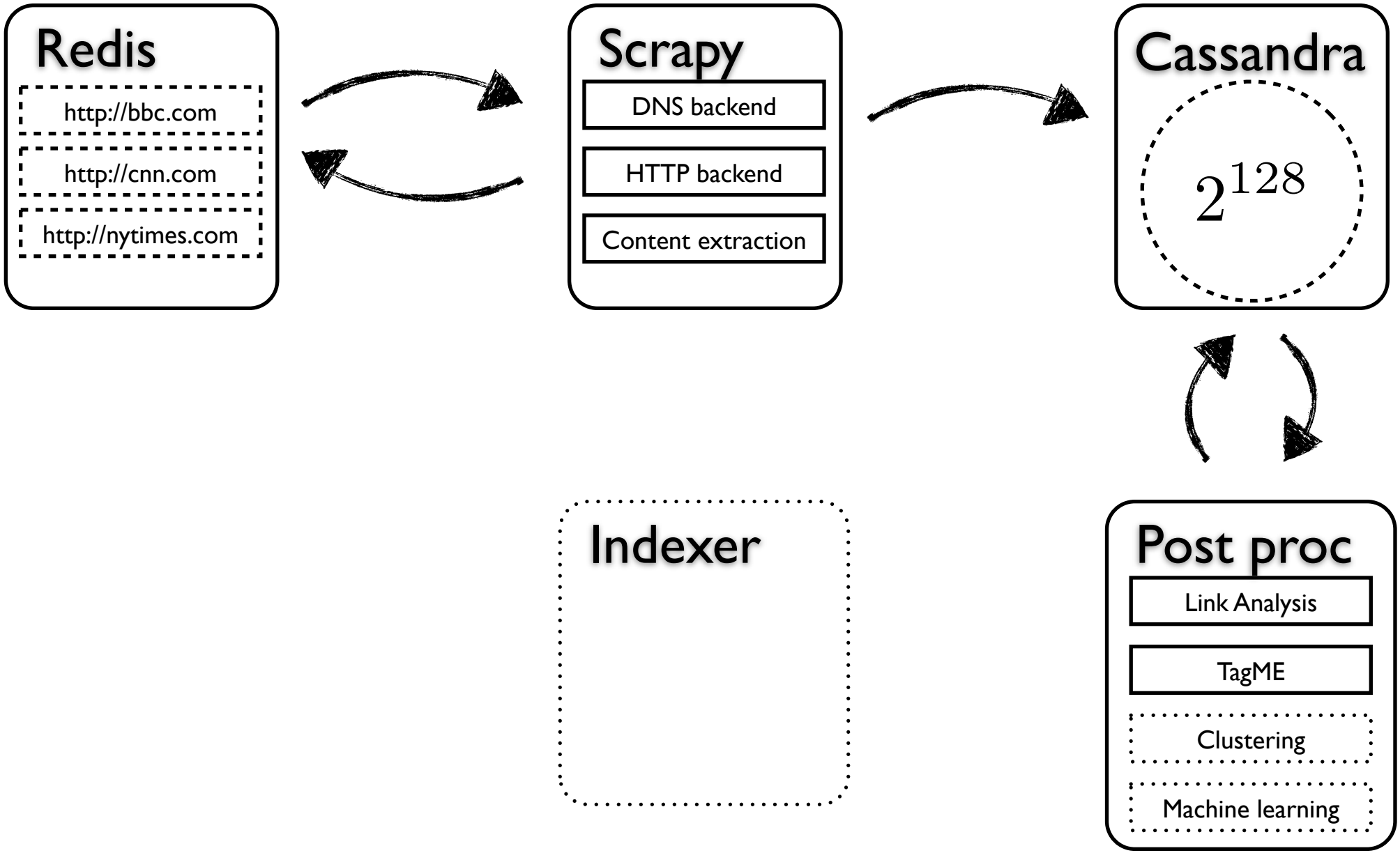
https://27esimaora.corriere.it/ ➡ https://27esimaora.corriere.it/articolo/bertolucci-e-la-scena-del..
https://27esimaora.corriere.it/ ➡ https://27esimaora.corriere.it/articolo/divorzio-allitaliana-ecco-per..
https://27esimaora.corriere.it/ ➡ https://27esimaora.corriere.it/articolo/donne-ingegnere-cinque-volte-meno-pag..
https://27esimaora.corriere.it/ ➡ https://27esimaora.corriere.it/articolo/donne-liberal-alla-conquista-del-sud..

↳ Spark cluster → Indexing phase

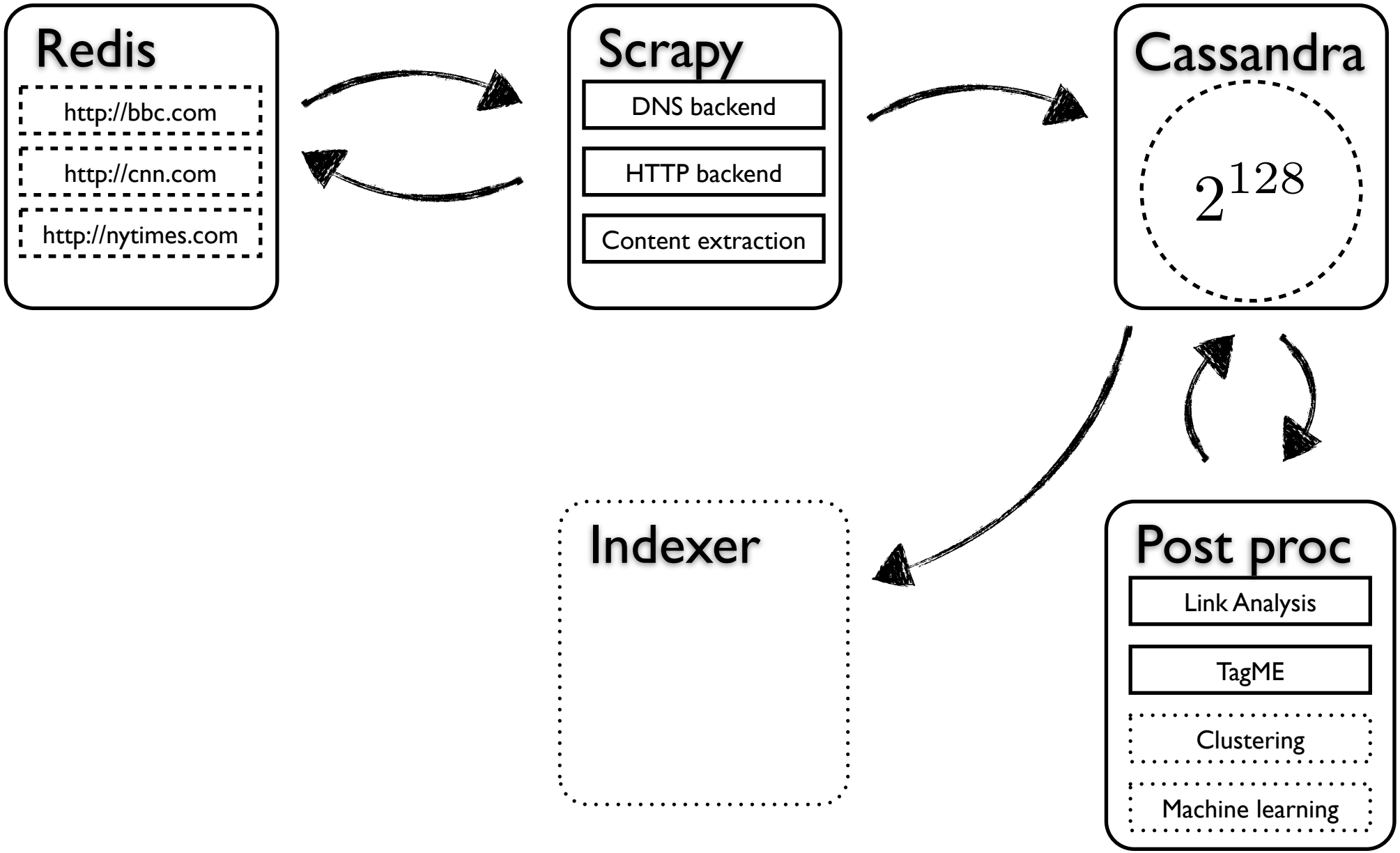
The big picture



The big picture



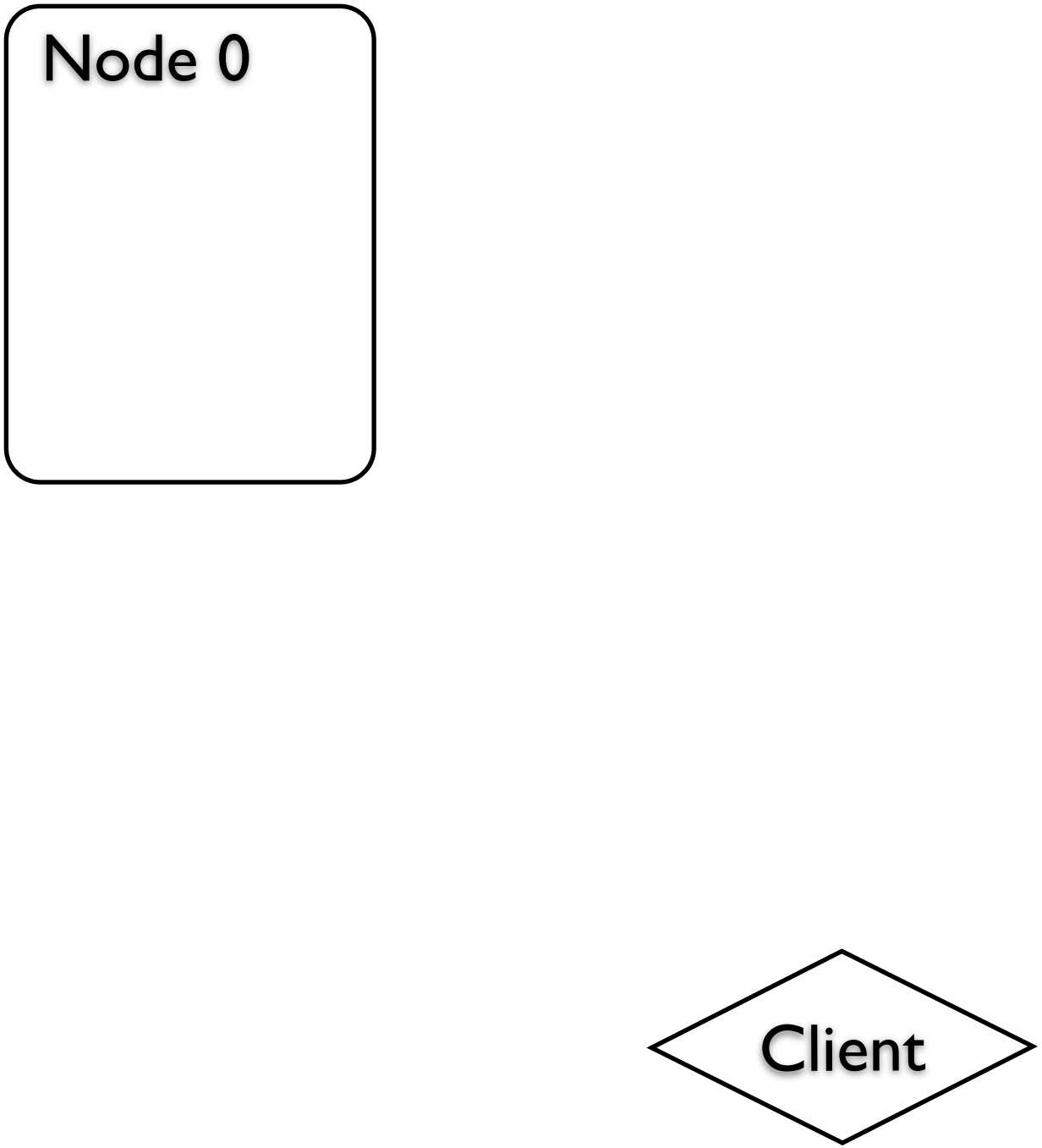
The big picture



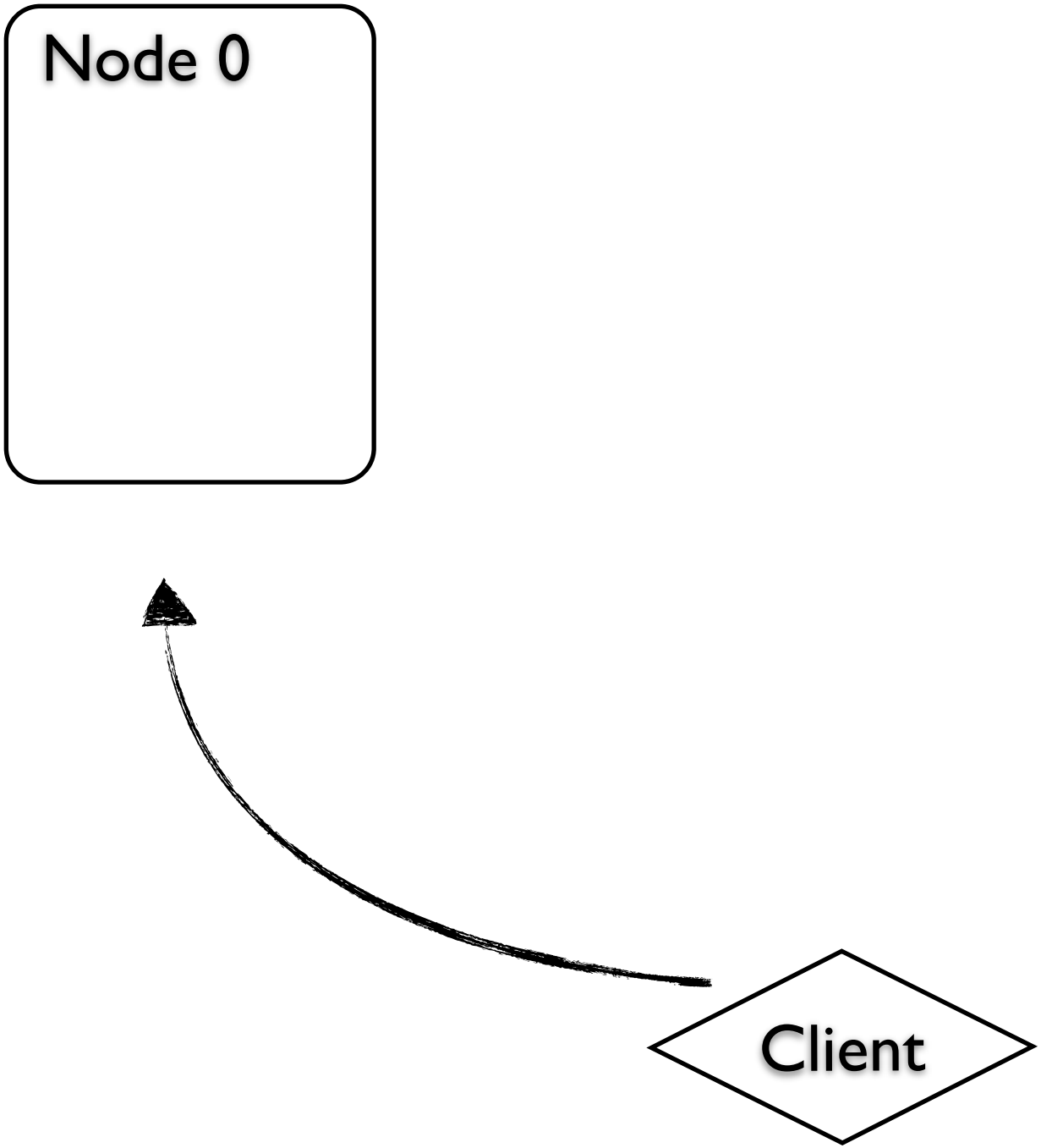
Indexing

- Why
  - Support term, phrase, prefix, regex.. queries
  - Suggest and autocompletion API
  - based on Lucene (Apache)
  - Plugins, REST api, easy
- Internals
  - Each index is sharded by documents
  - Each shard can have zero or more replicas
  - Scaling by having more shards

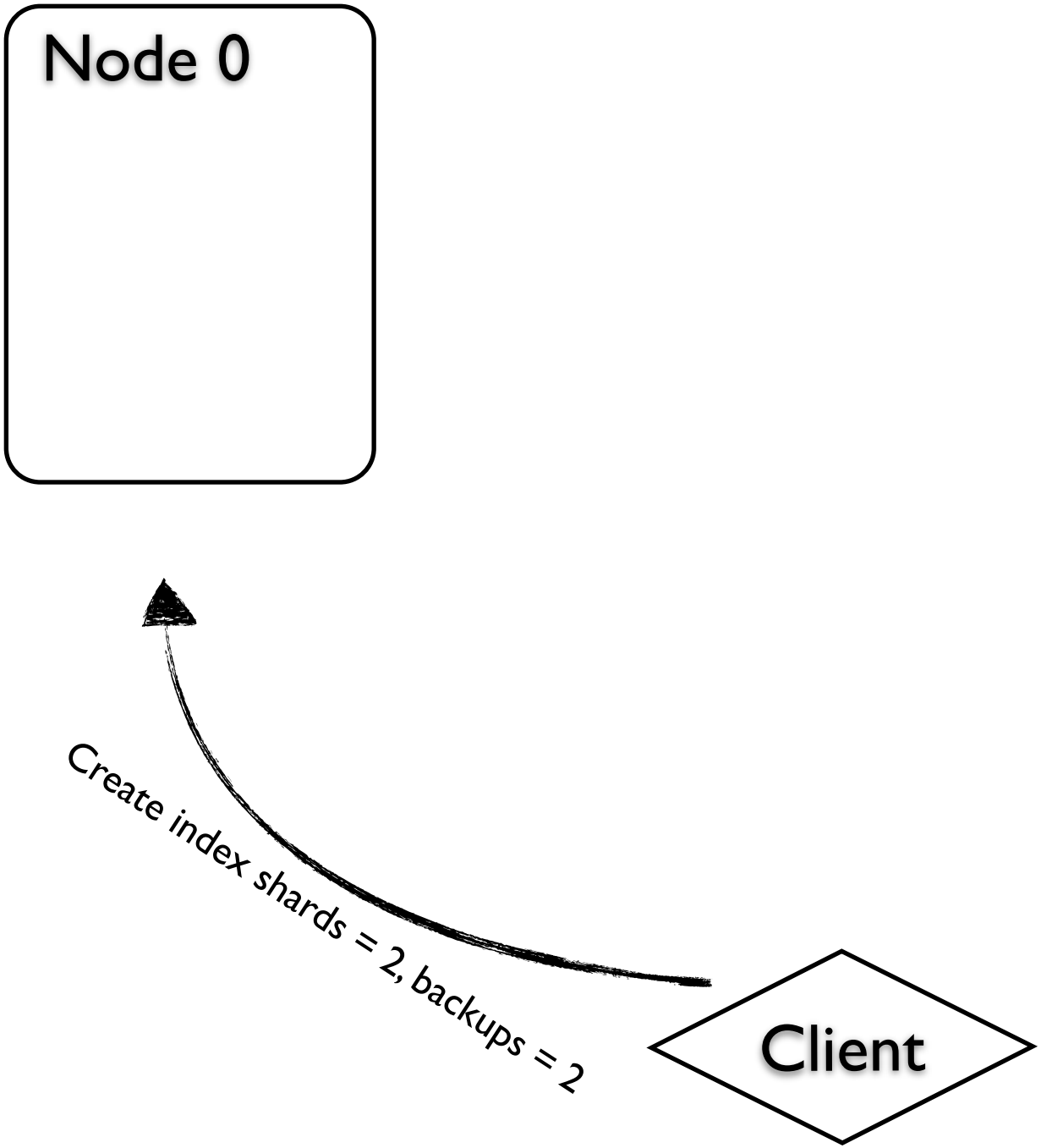
Distributed Elasticsearch



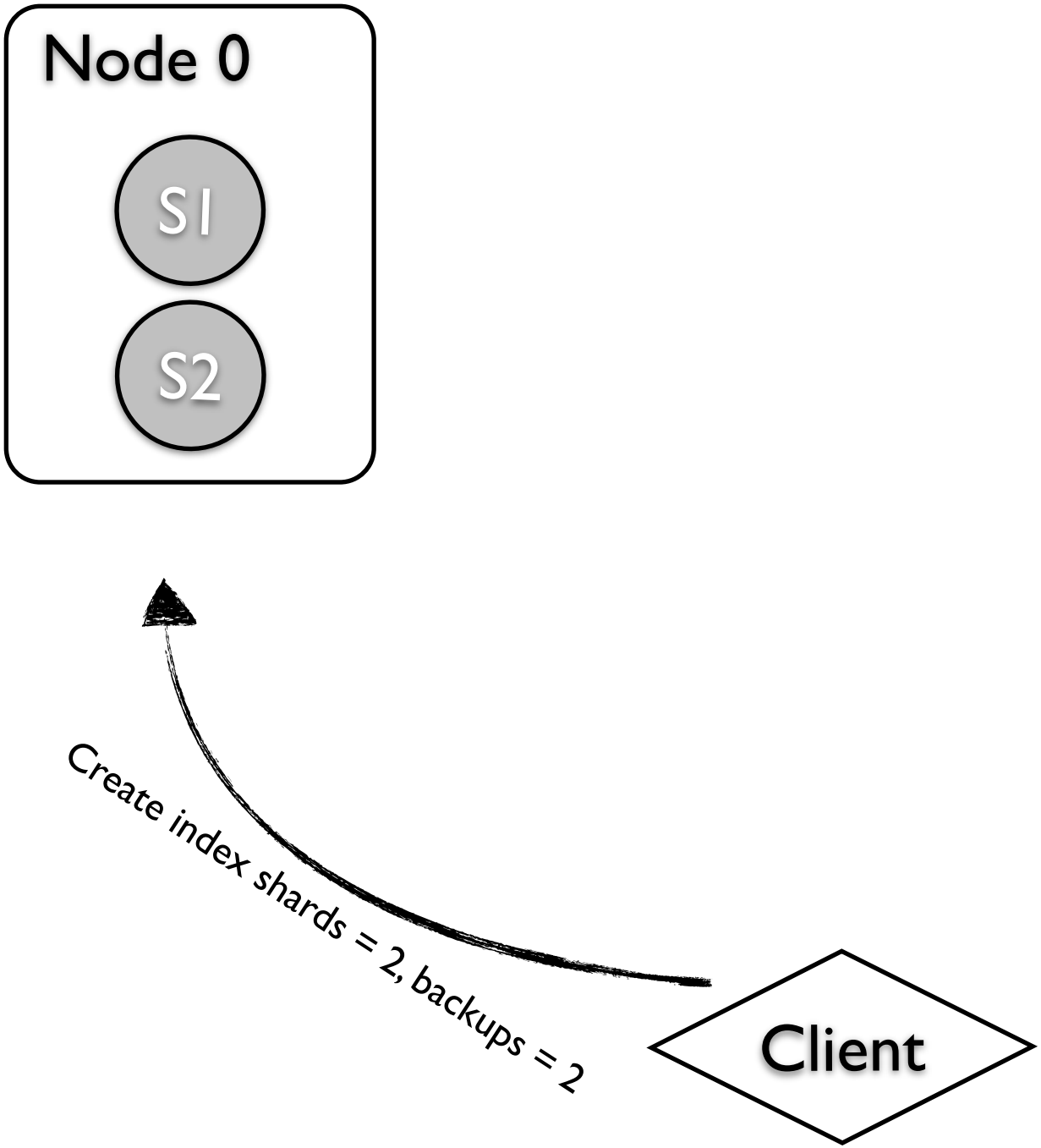
Distributed Elasticsearch



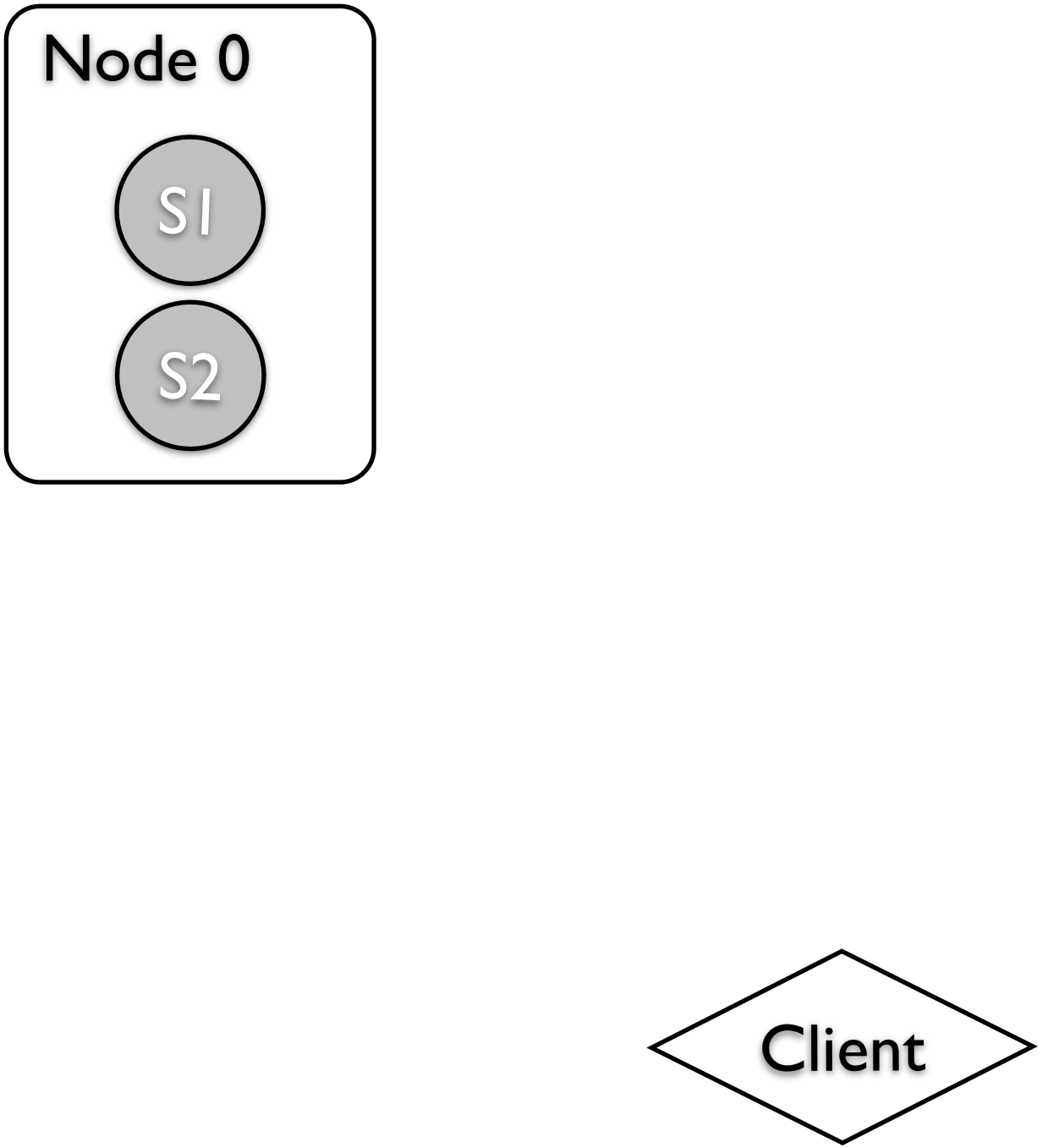
Distributed Elasticsearch



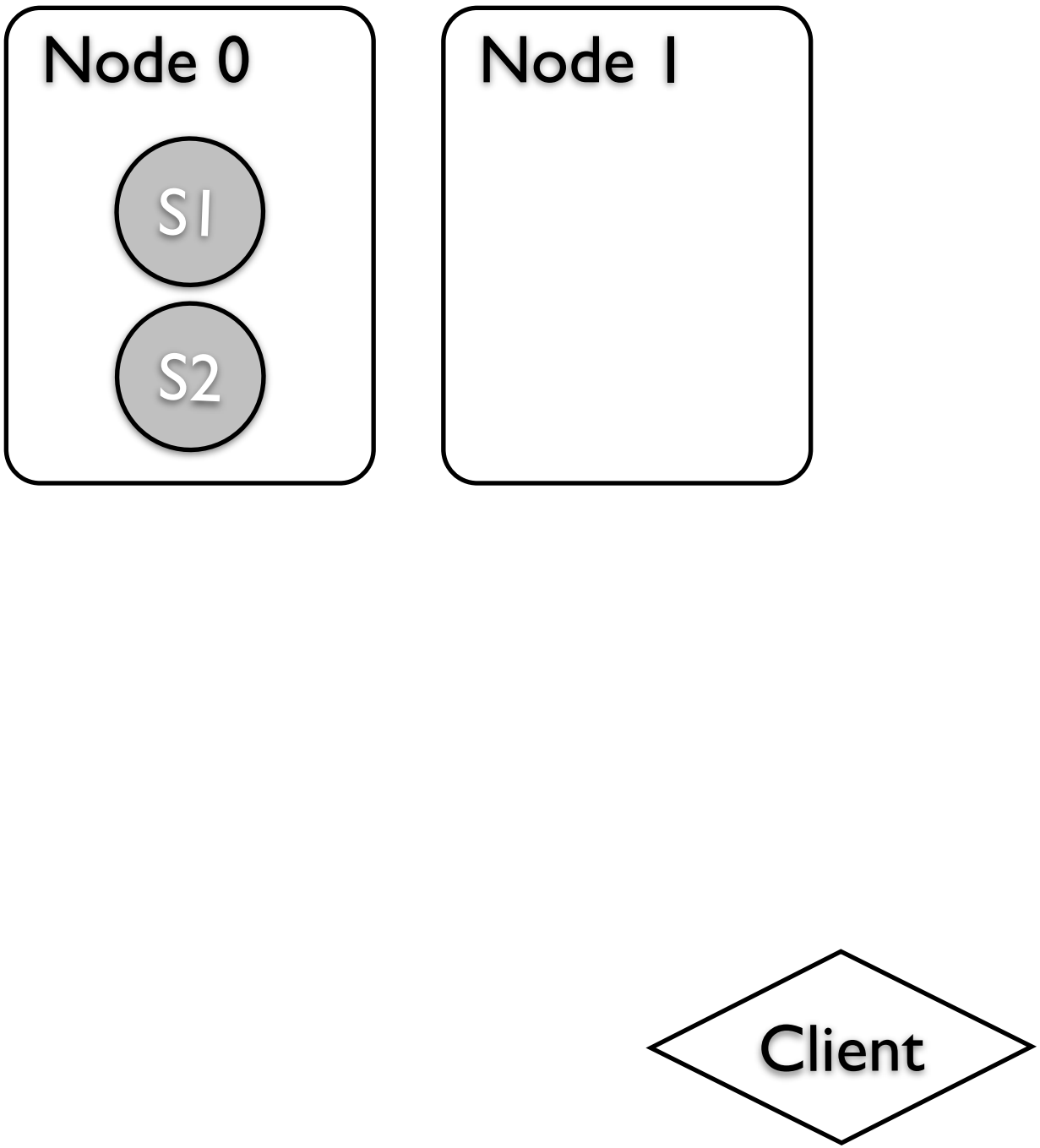
Distributed Elasticsearch



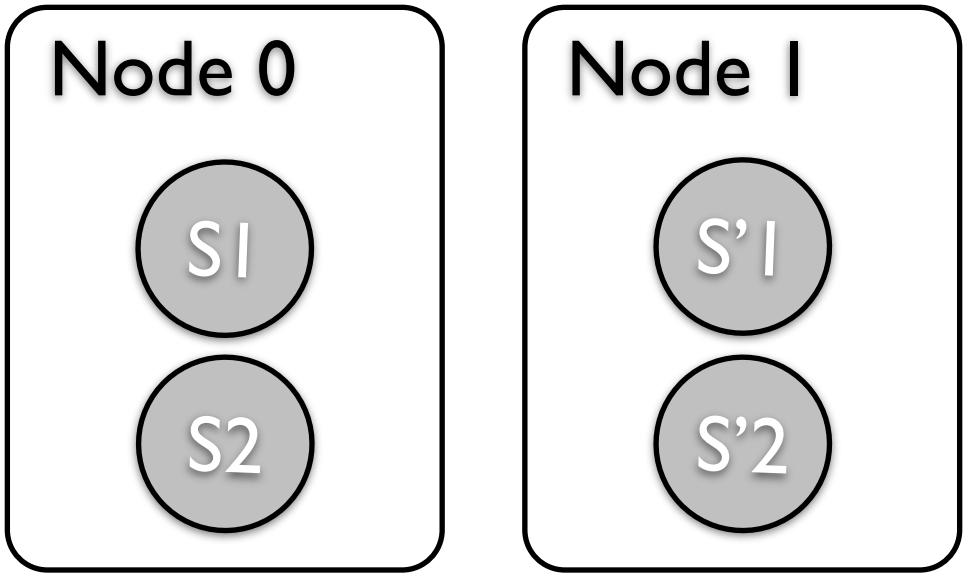
Distributed Elasticsearch



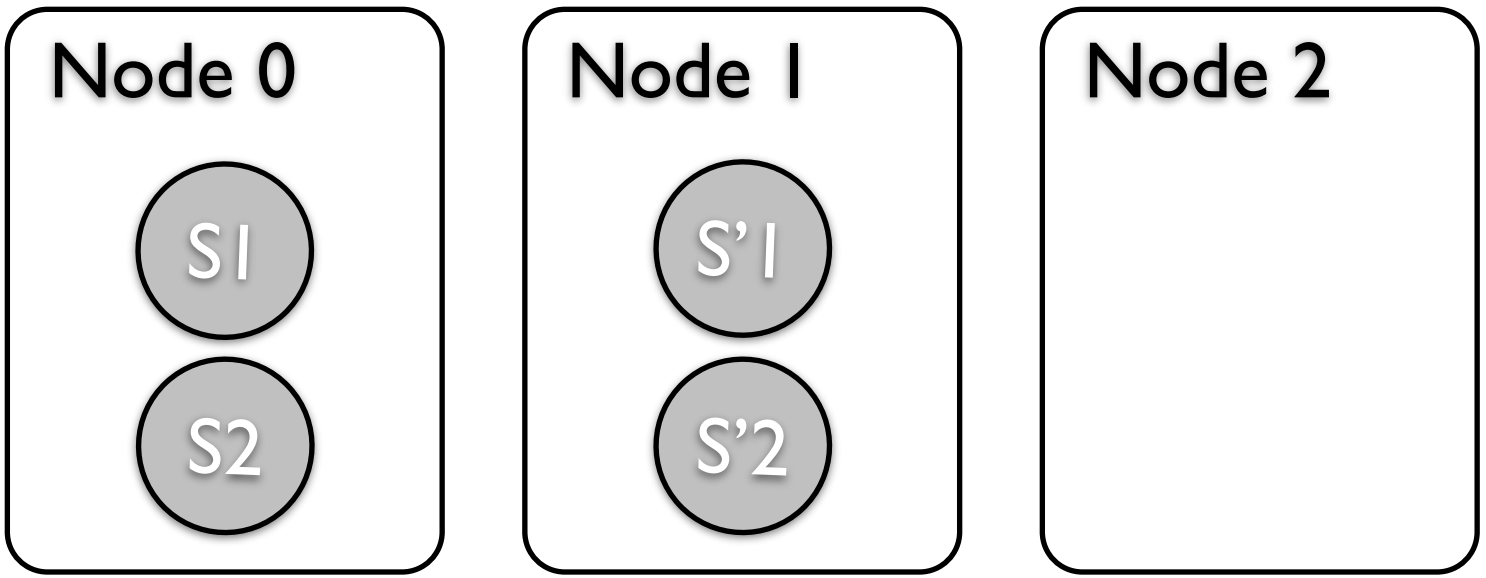
Distributed Elasticsearch



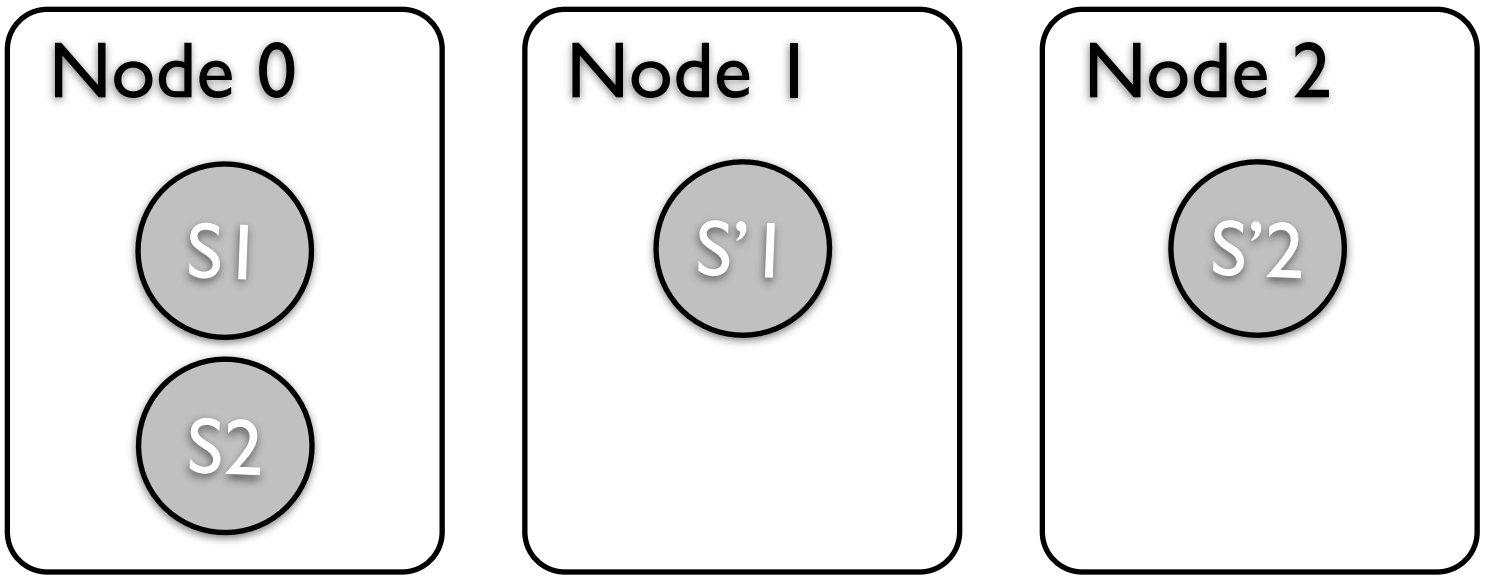
Distributed Elasticsearch



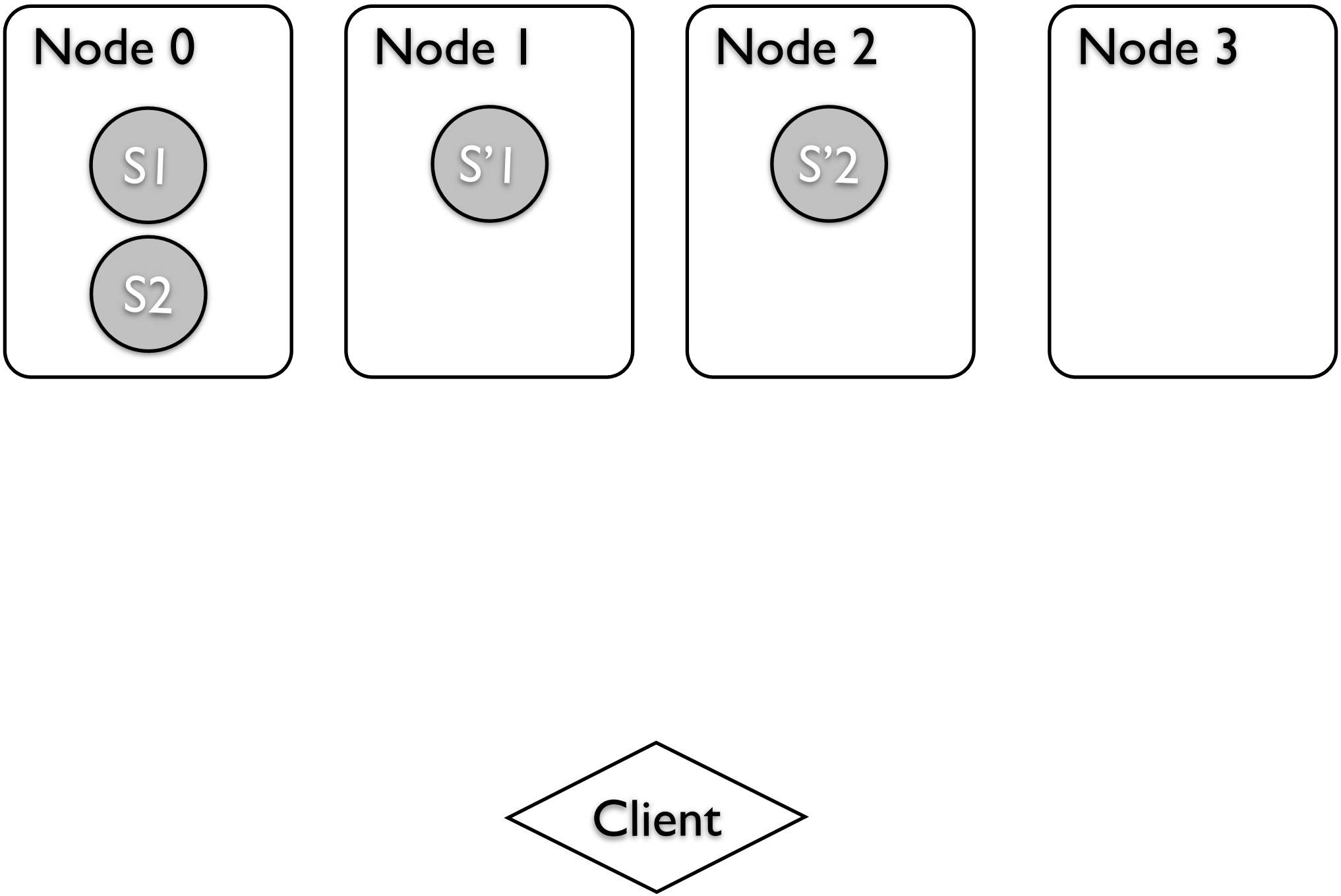
Distributed Elasticsearch



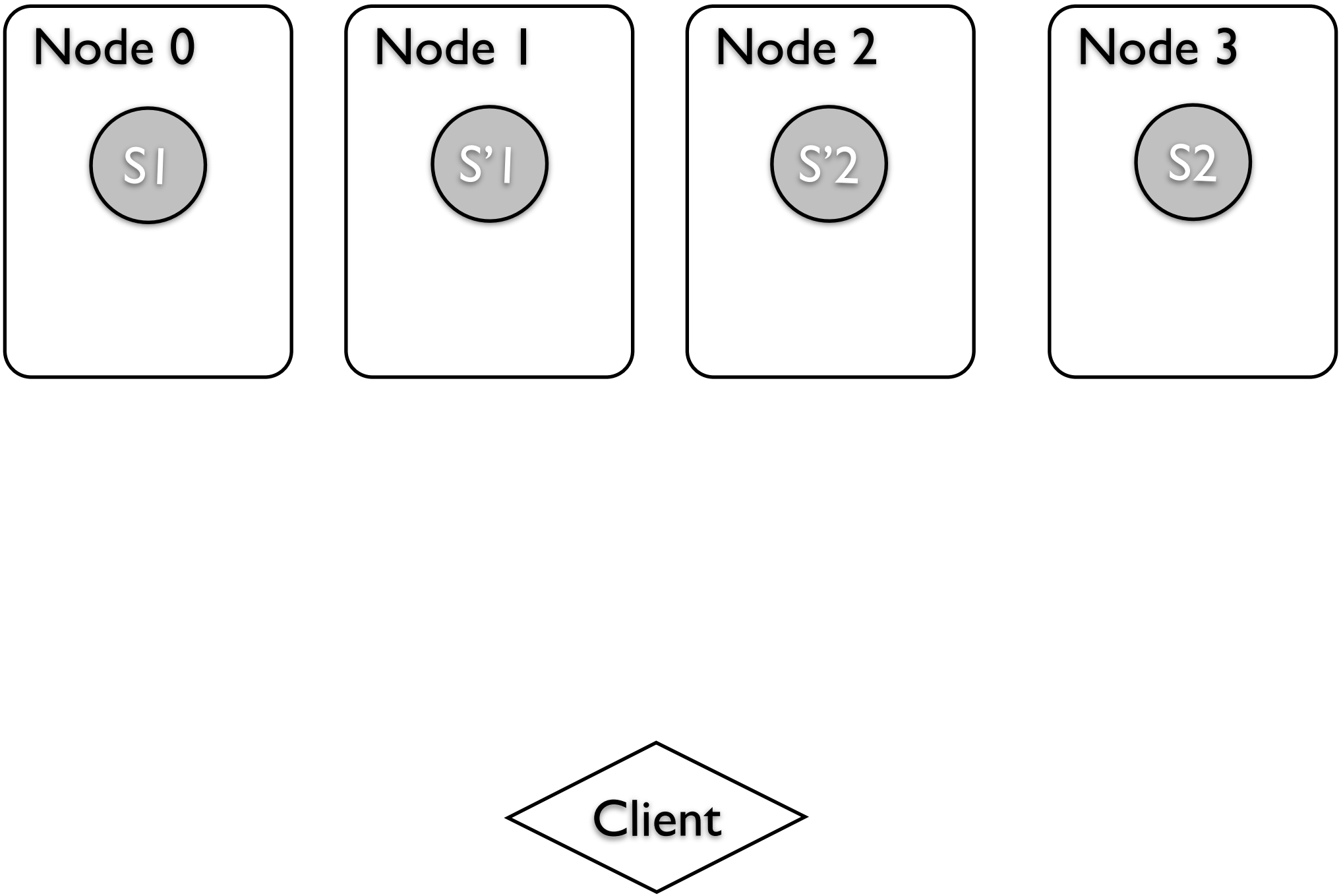
Distributed Elasticsearch



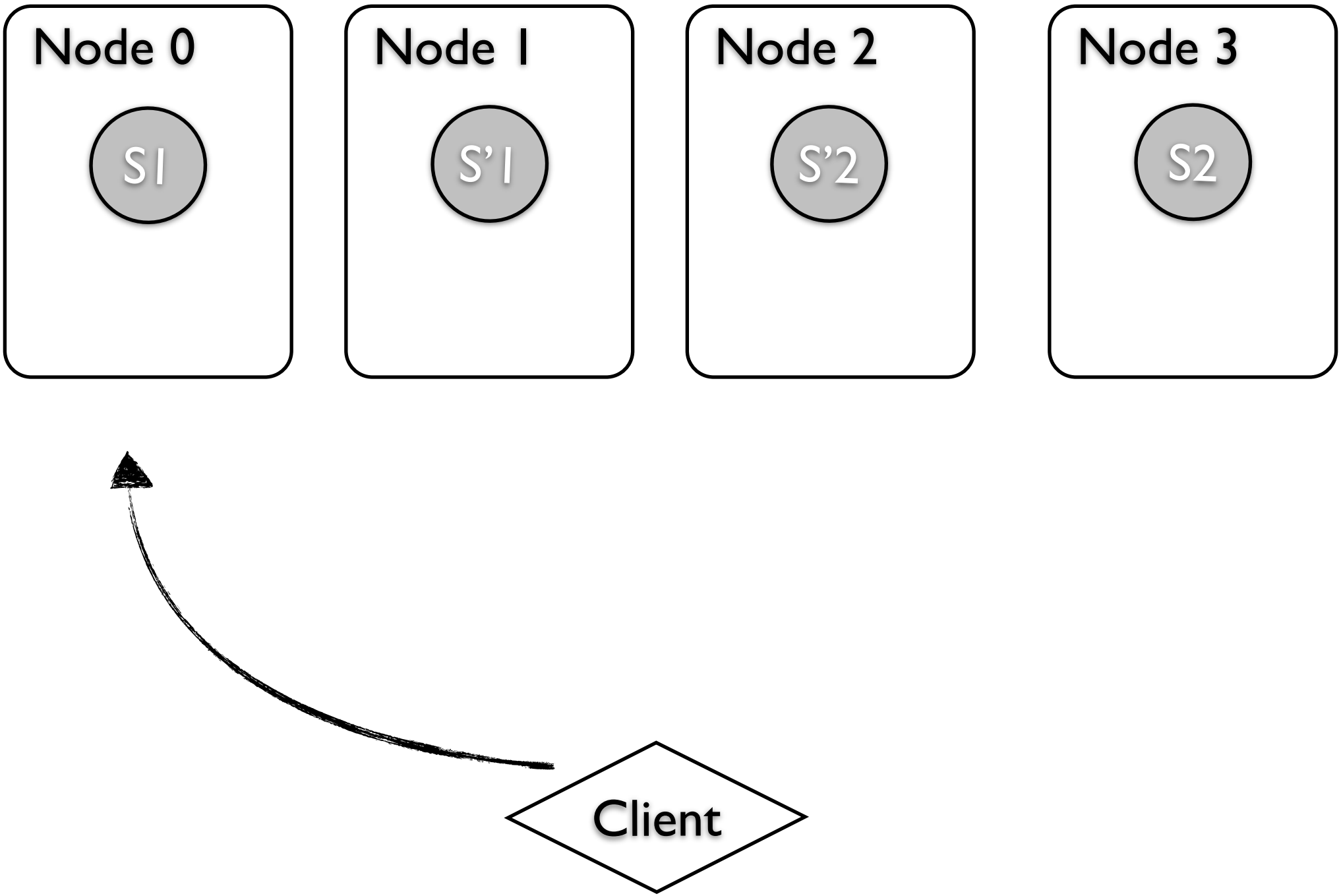
Distributed Elasticsearch



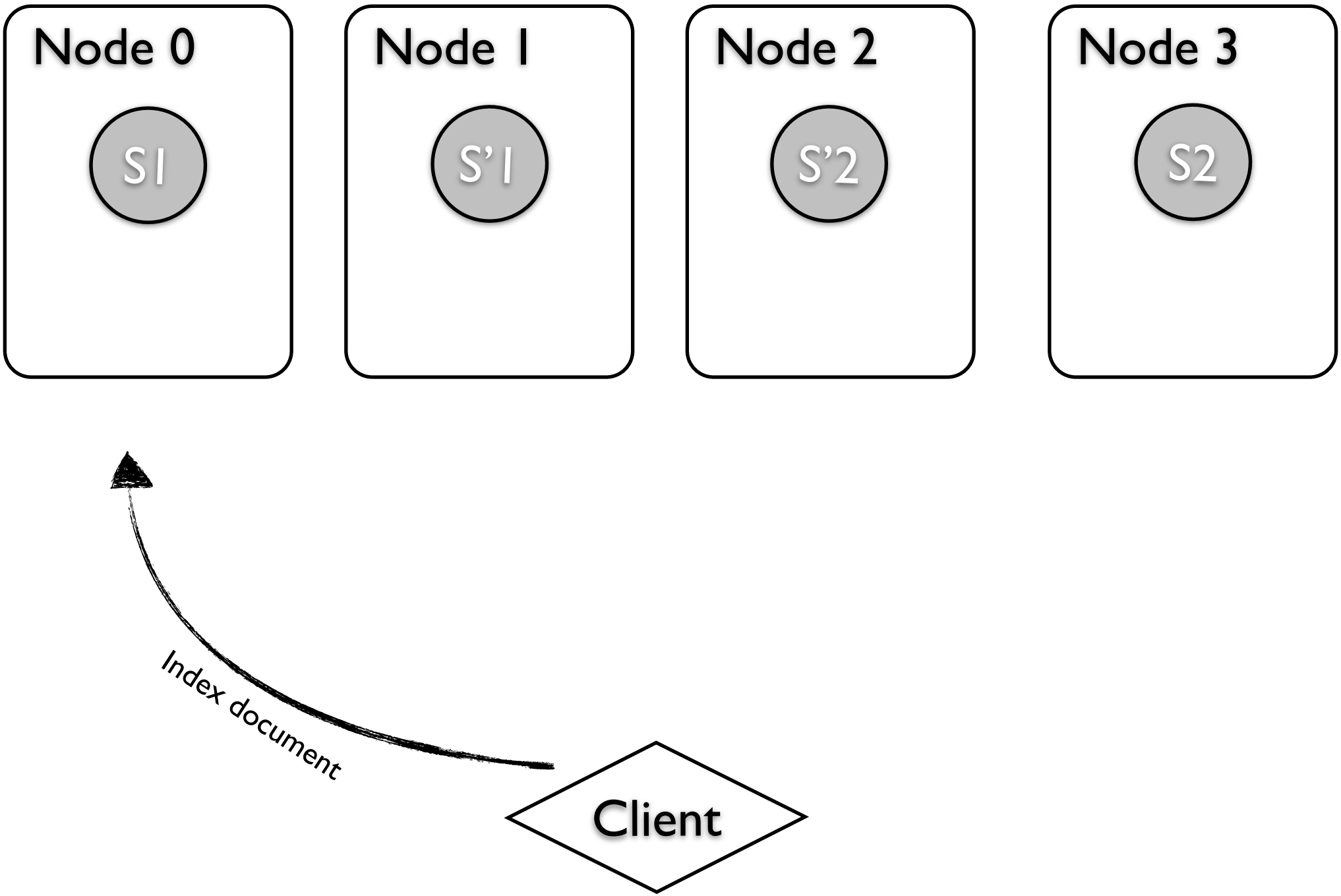
Distributed Elasticsearch



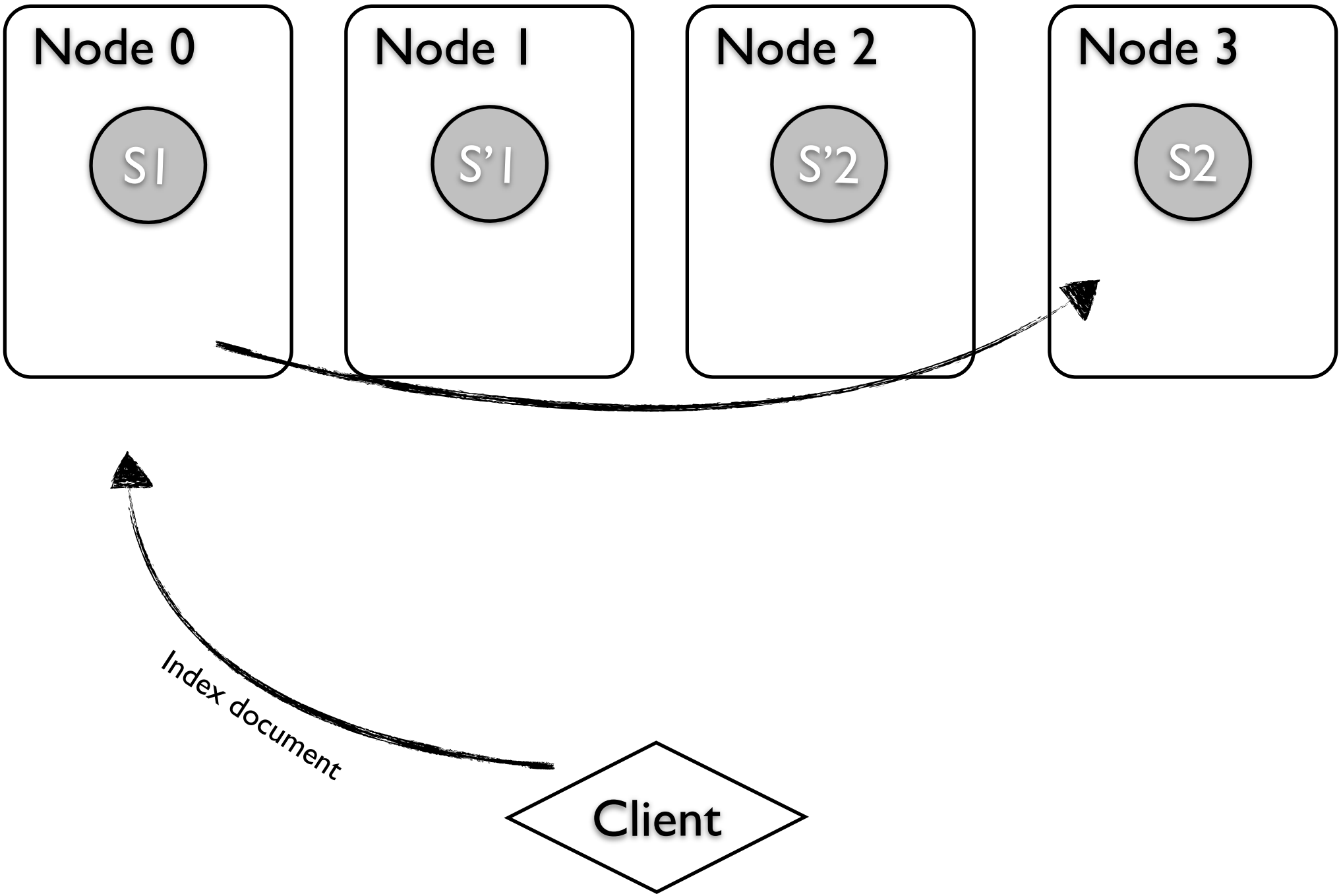
Distributed Elasticsearch



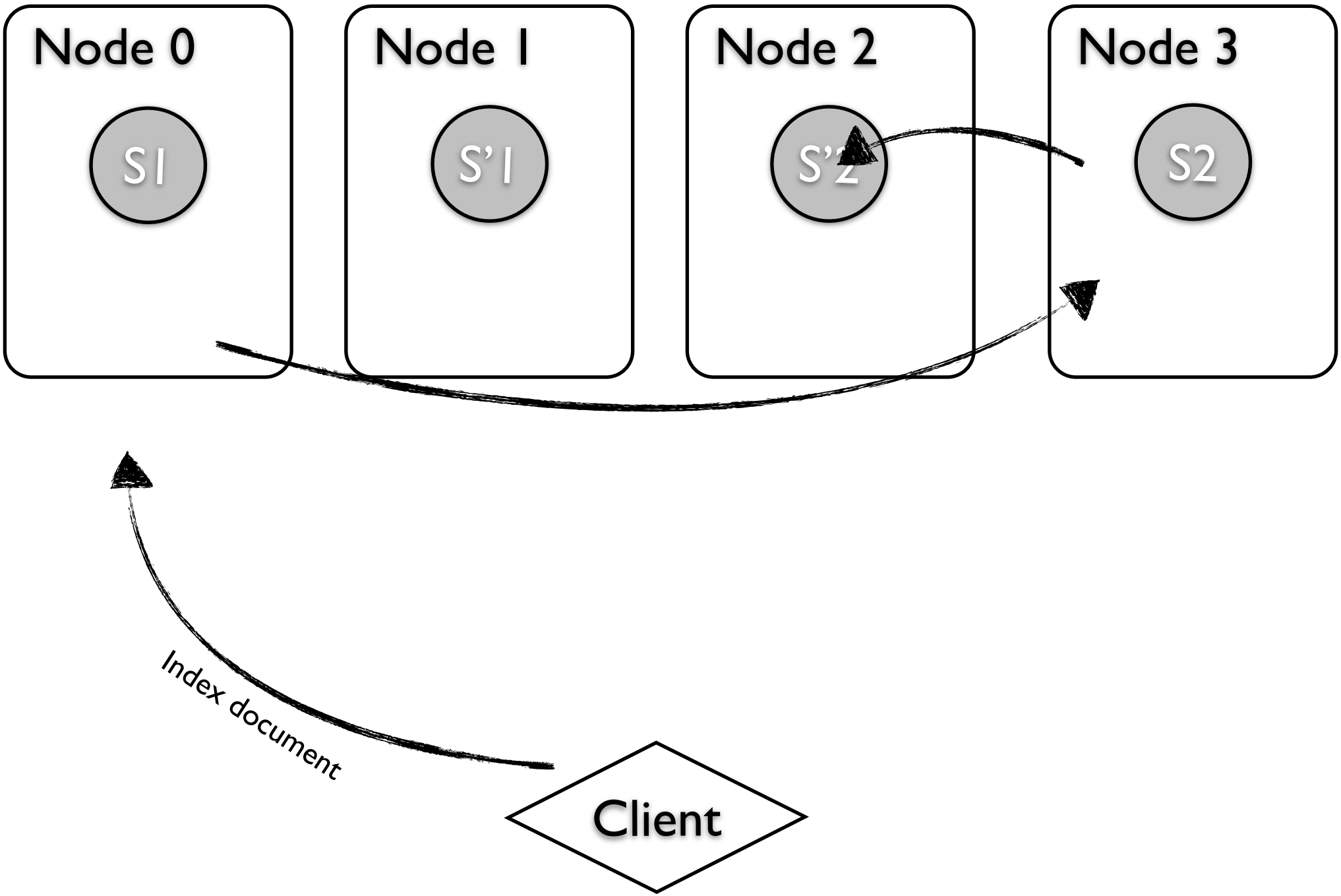
Distributed Elasticsearch



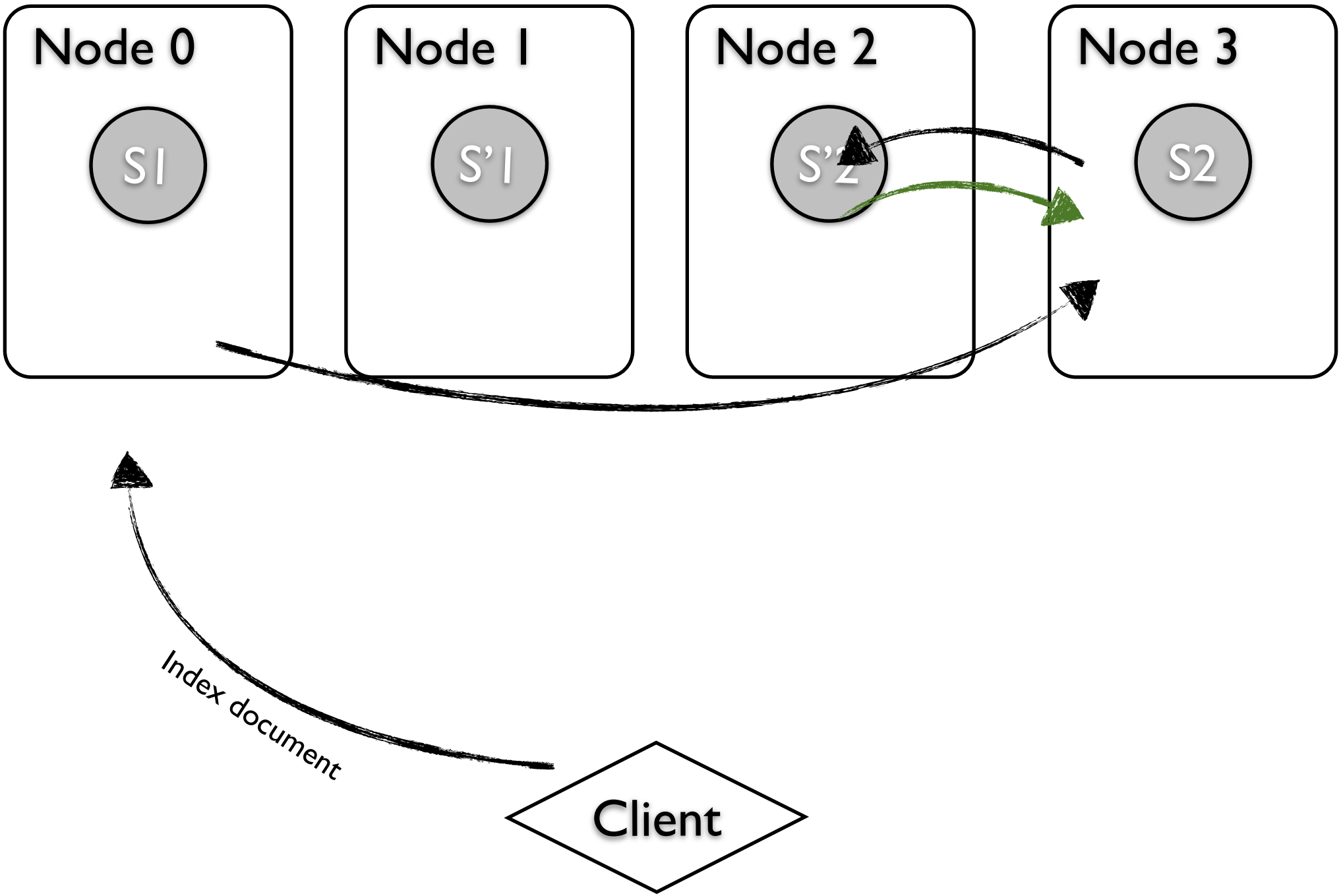
Distributed Elasticsearch



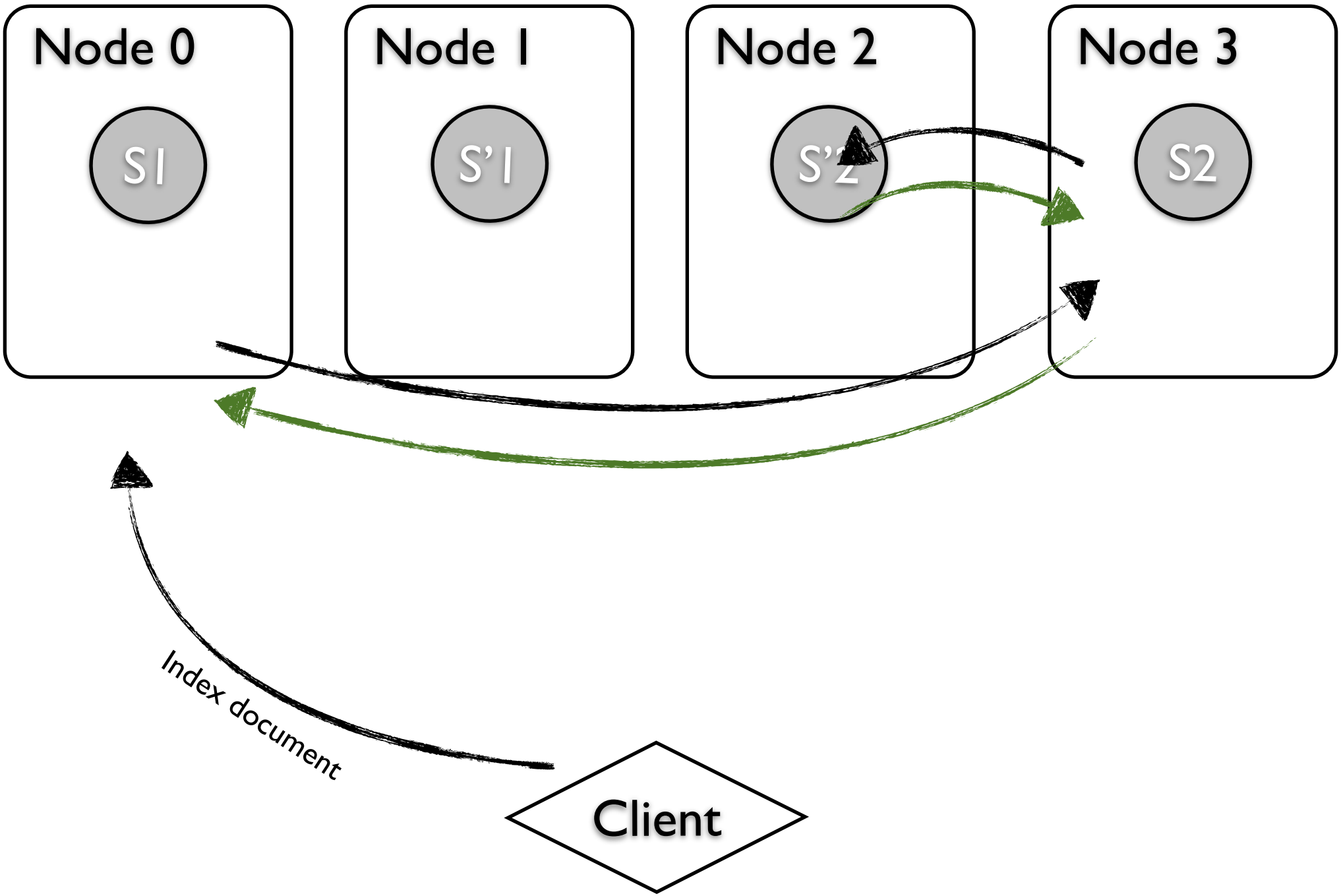
Distributed Elasticsearch



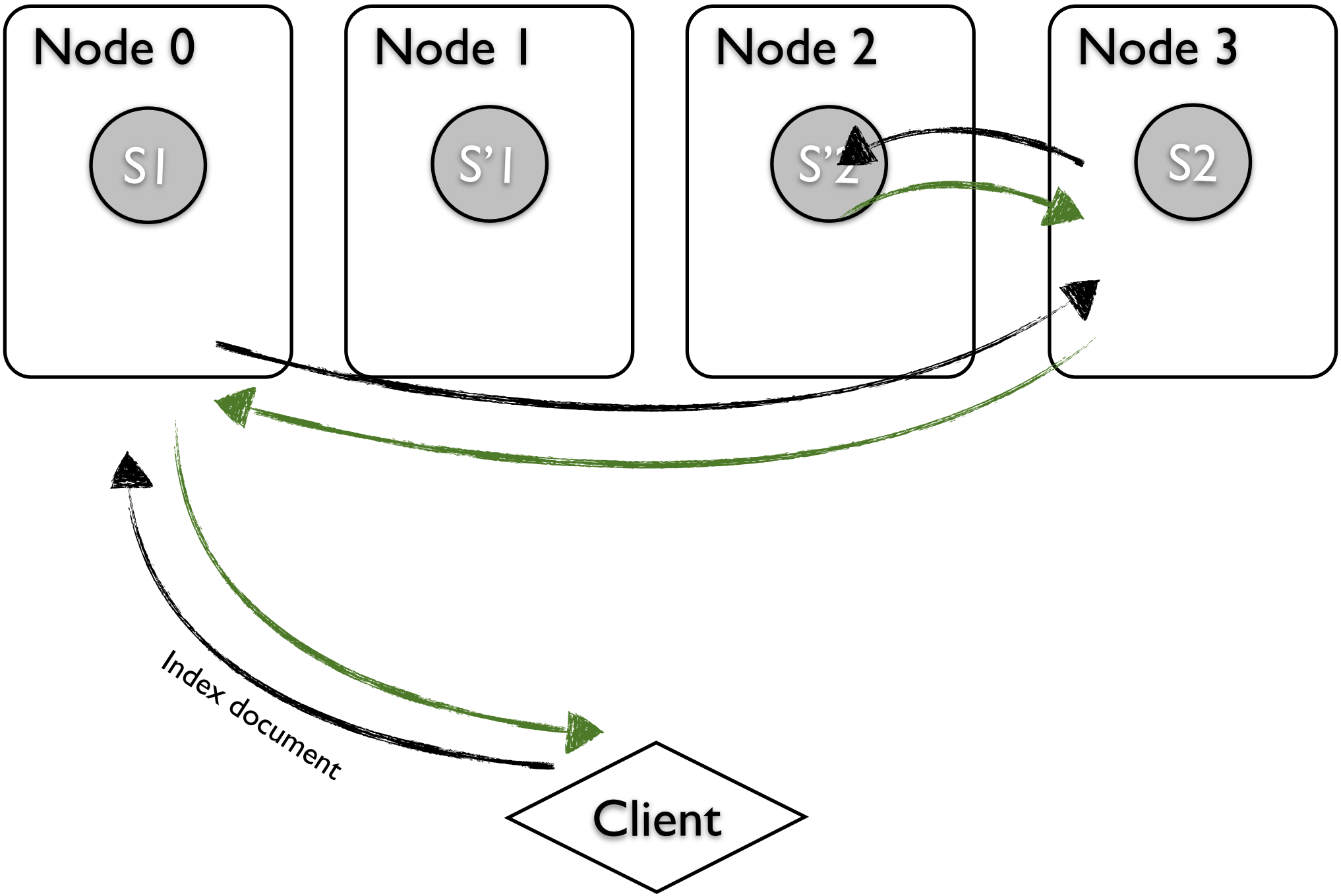
Distributed Elasticsearch



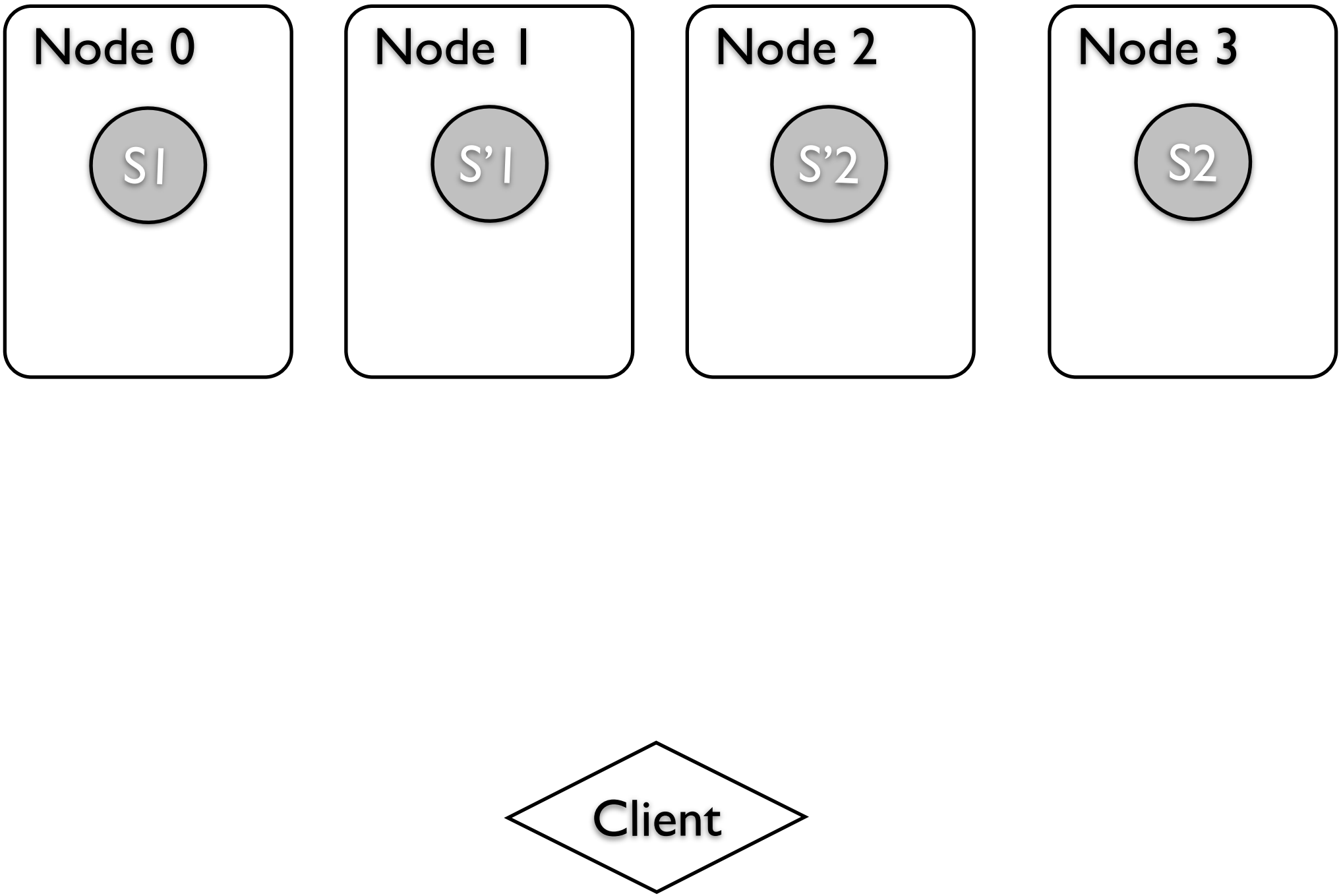
Distributed Elasticsearch



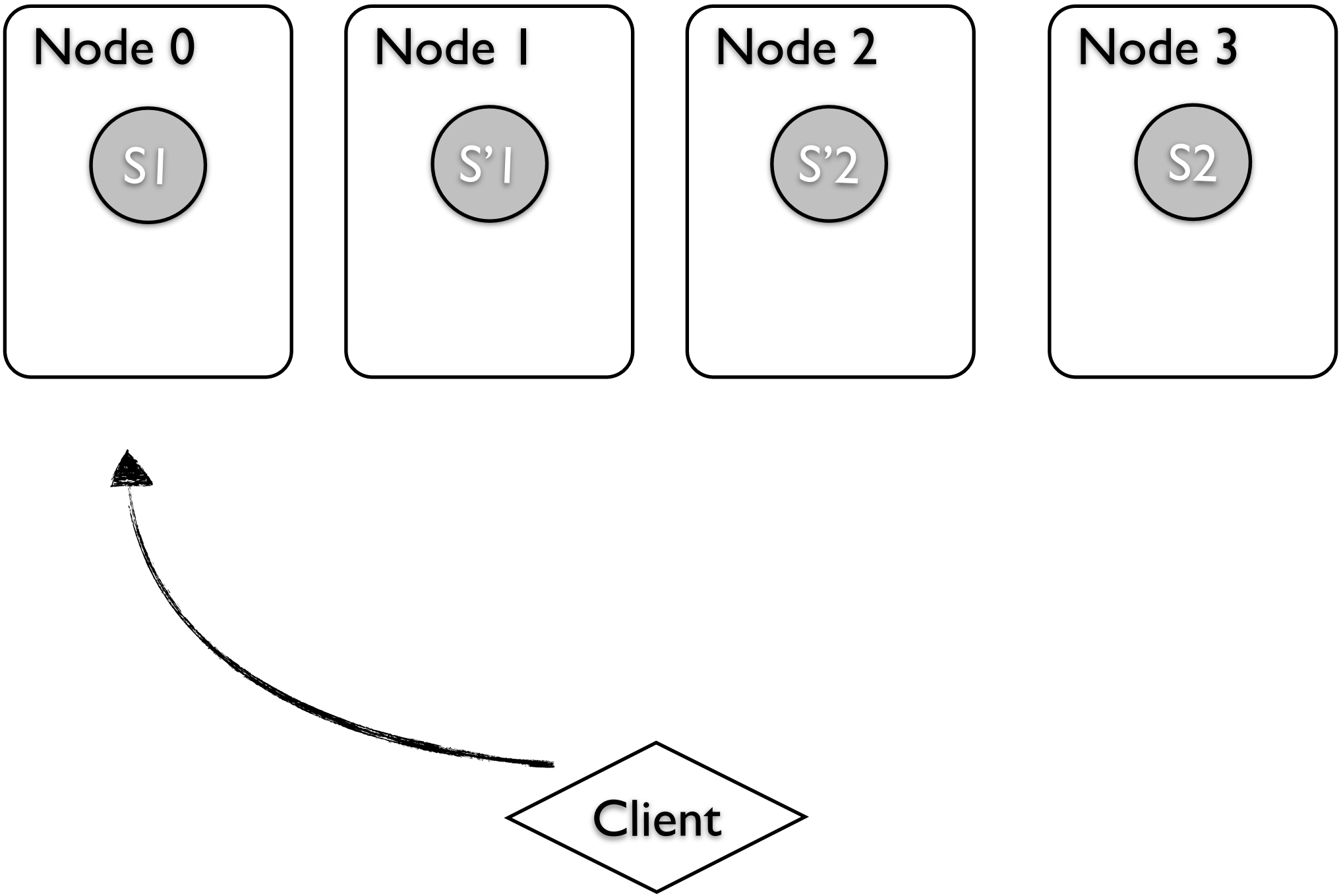
Distributed Elasticsearch



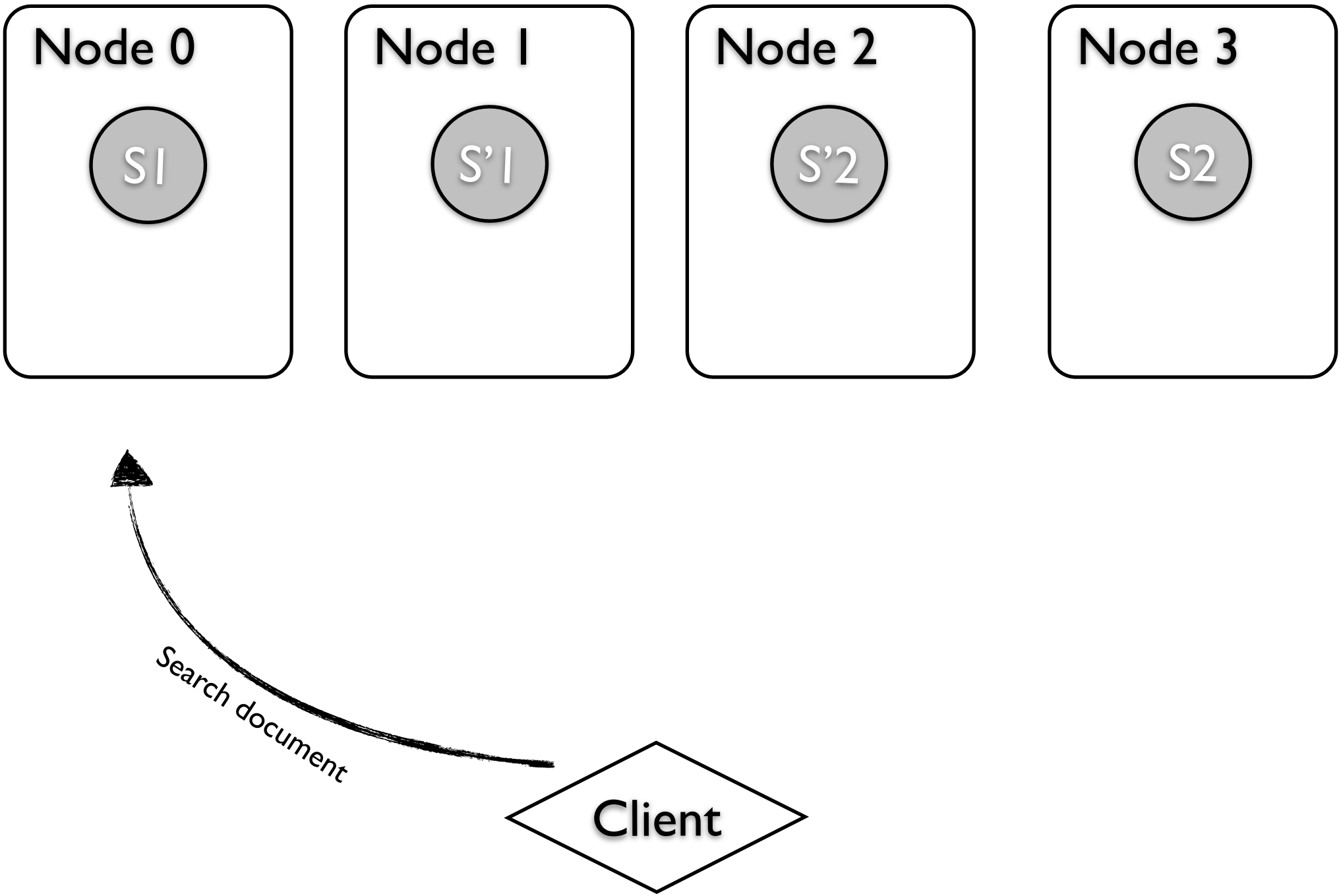
Distributed Elasticsearch



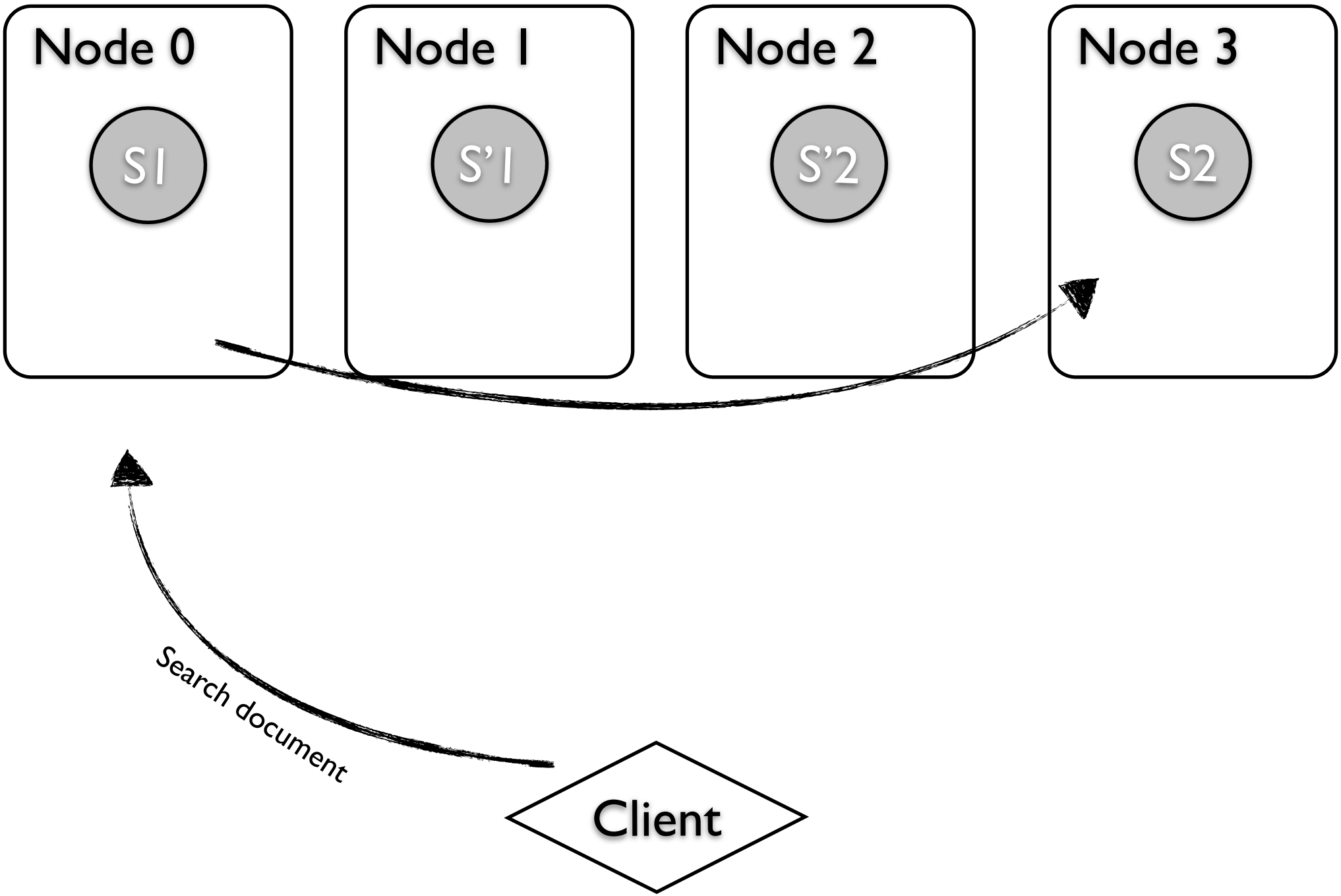
Distributed Elasticsearch



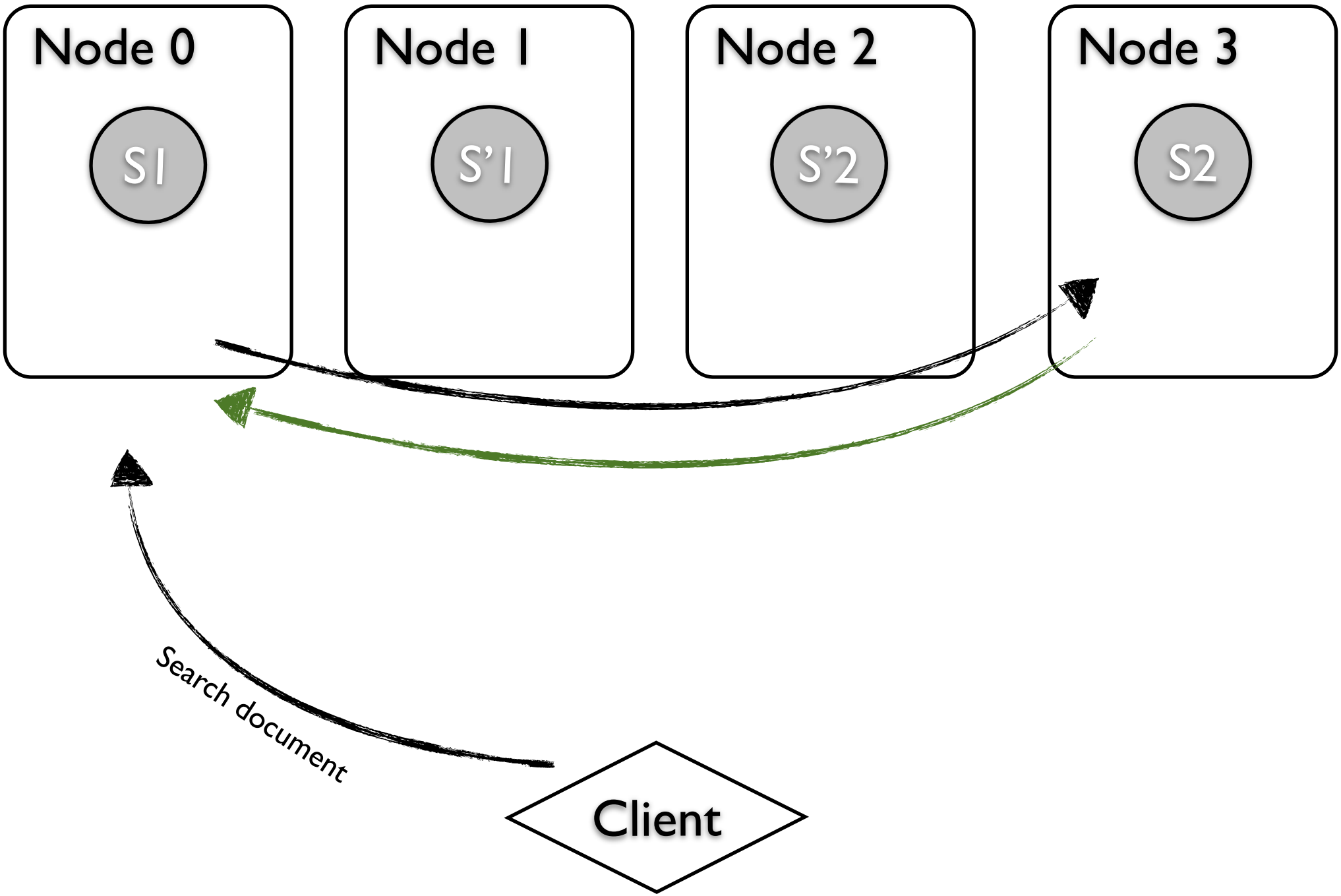
Distributed Elasticsearch



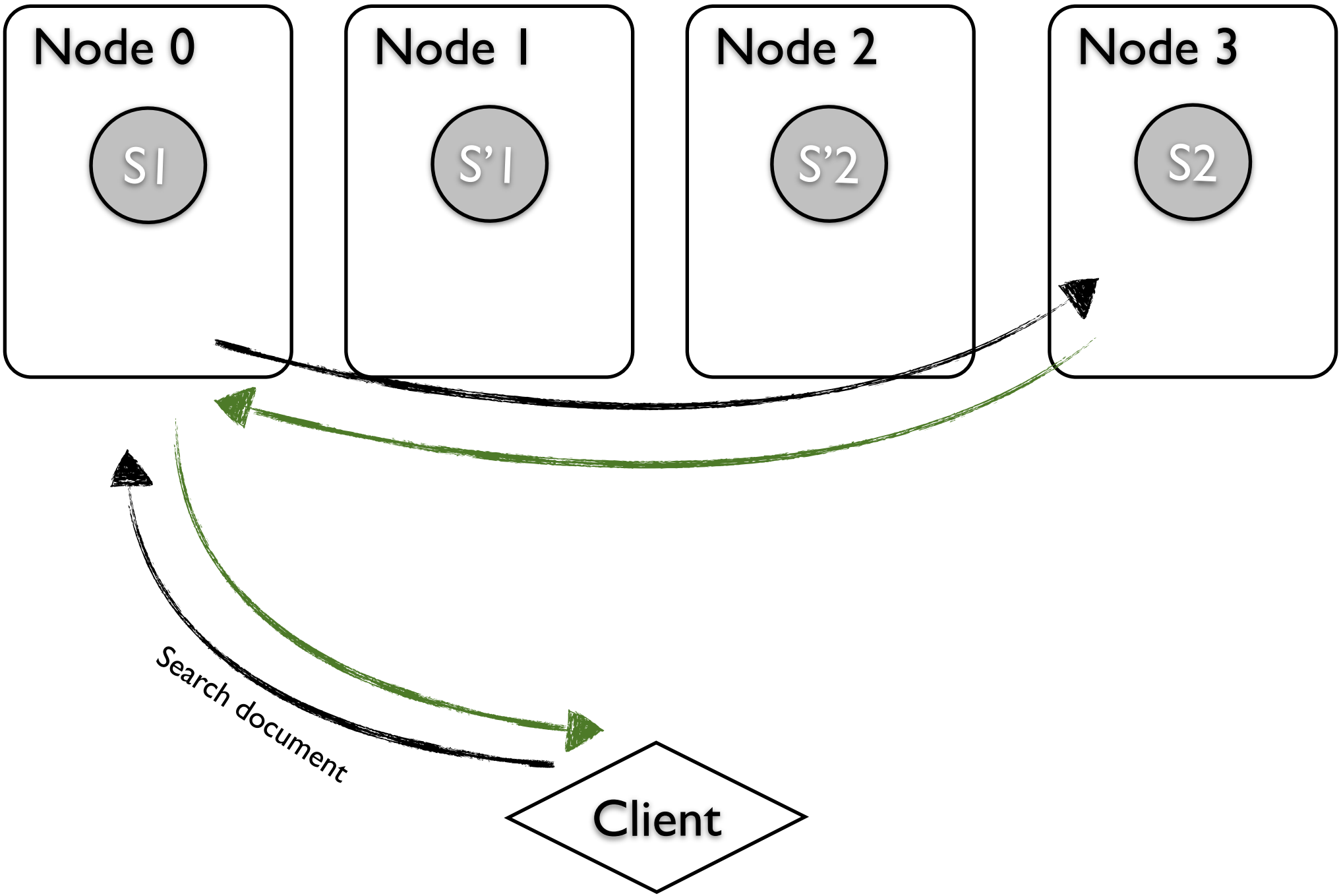
Distributed Elasticsearch



Distributed Elasticsearch

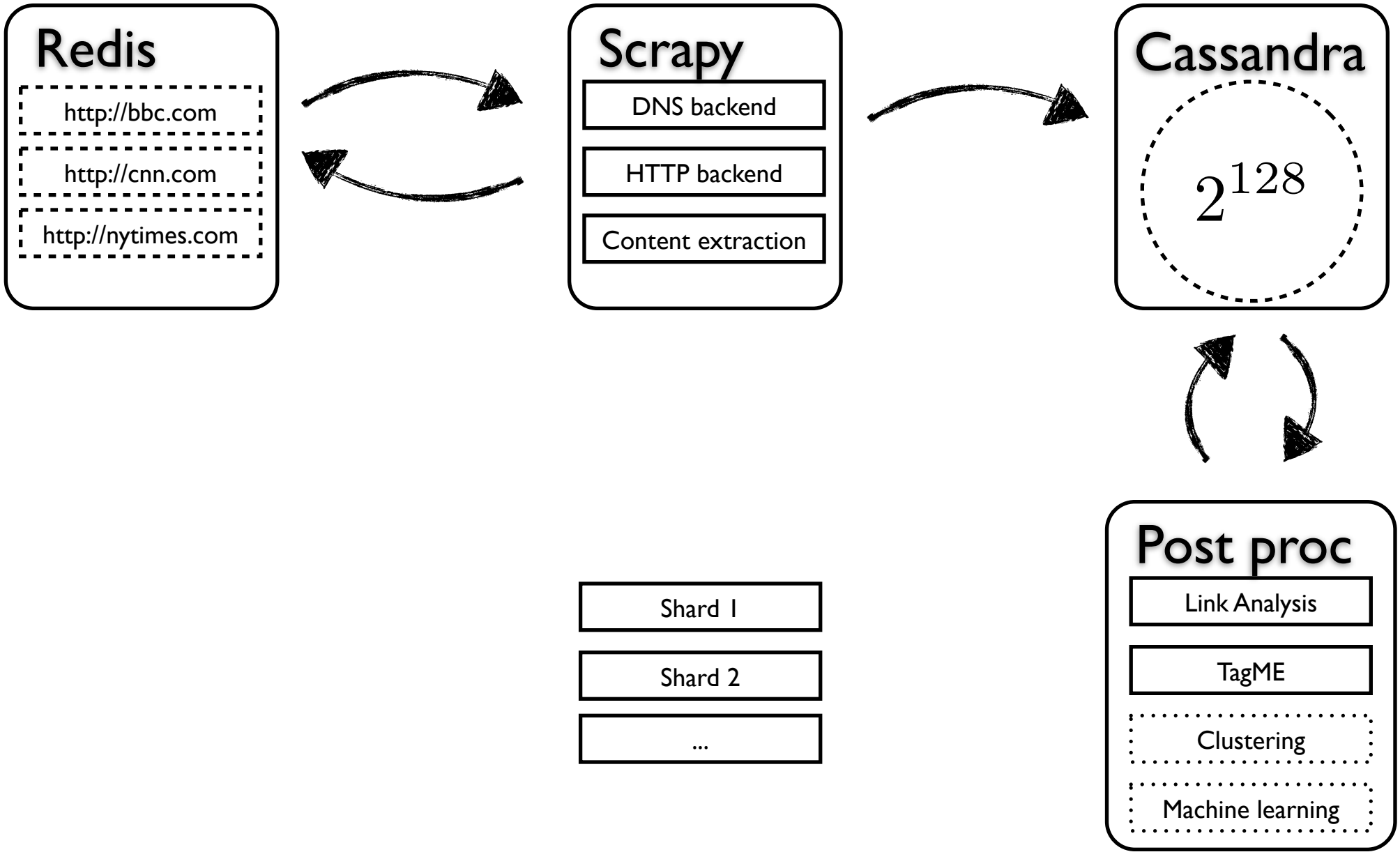


Distributed Elasticsearch

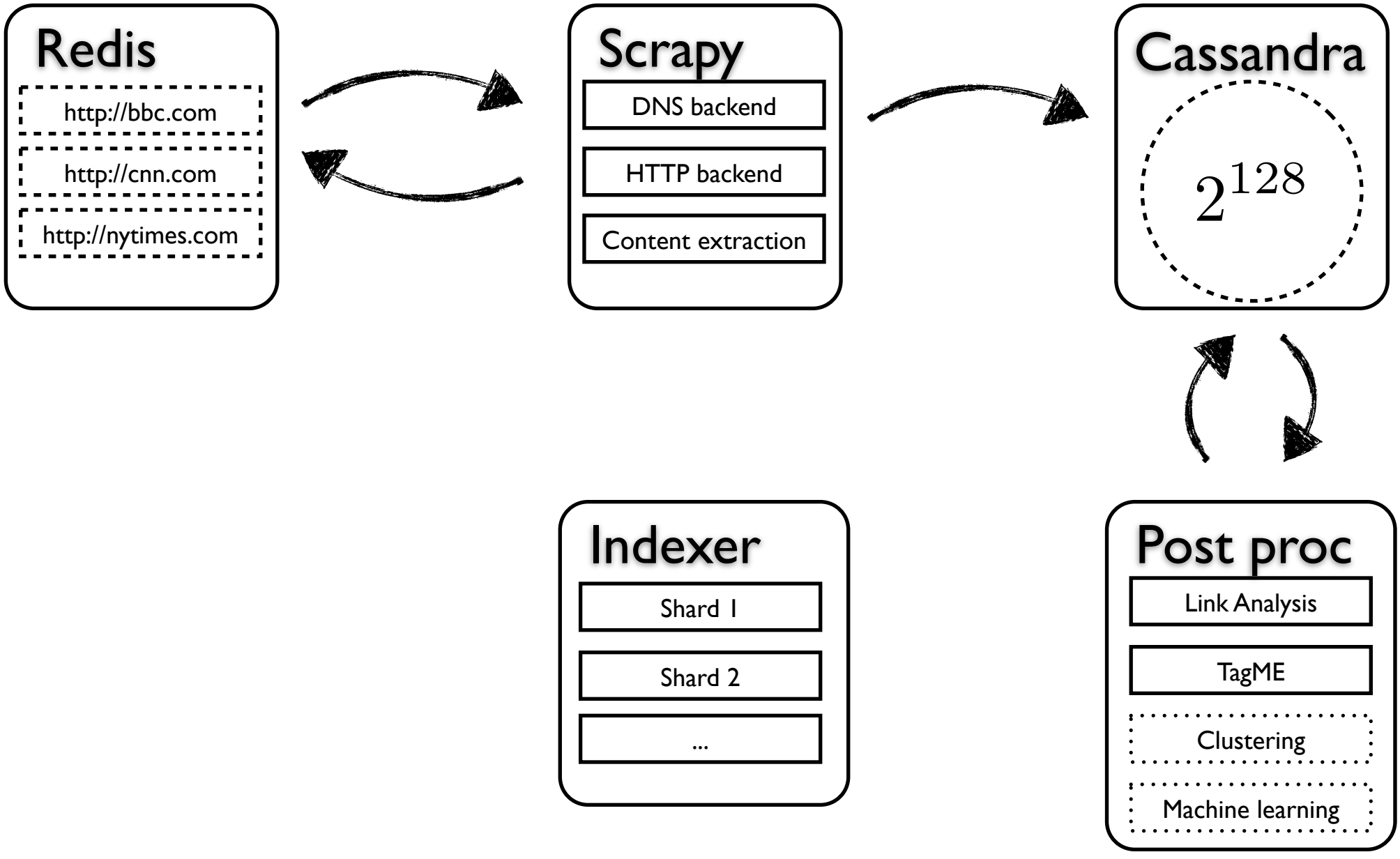


Demo

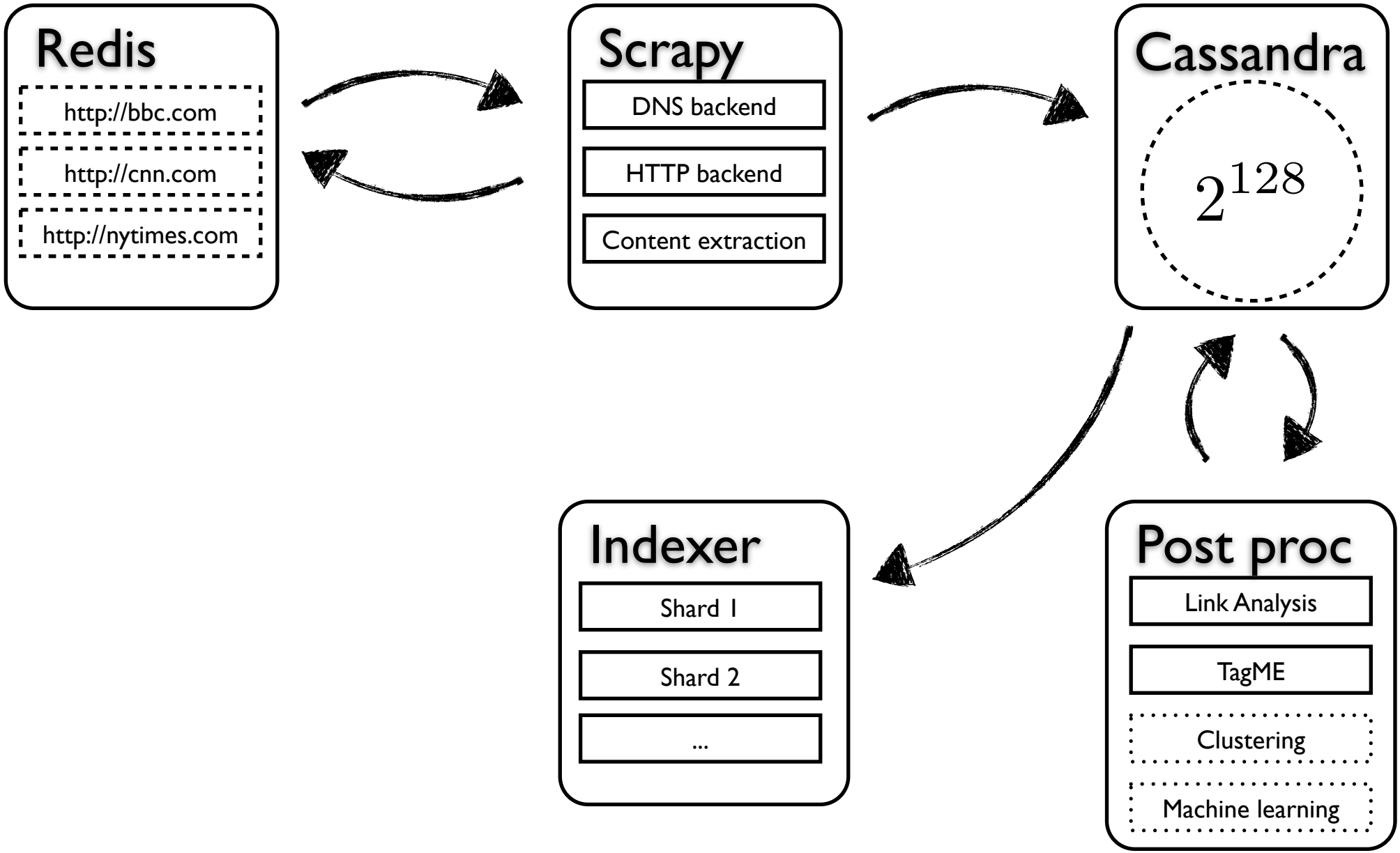
Backend


Redis
http://bbc.com
http://cnn.com
http://nytimes.com
Scrapy
DNS backend
HTTP backend
Content extraction
Cassandra
2^{128}
Shard 1
Shard 2
...
Post proc
Link Analysis
TagME
Clustering
Machine learning

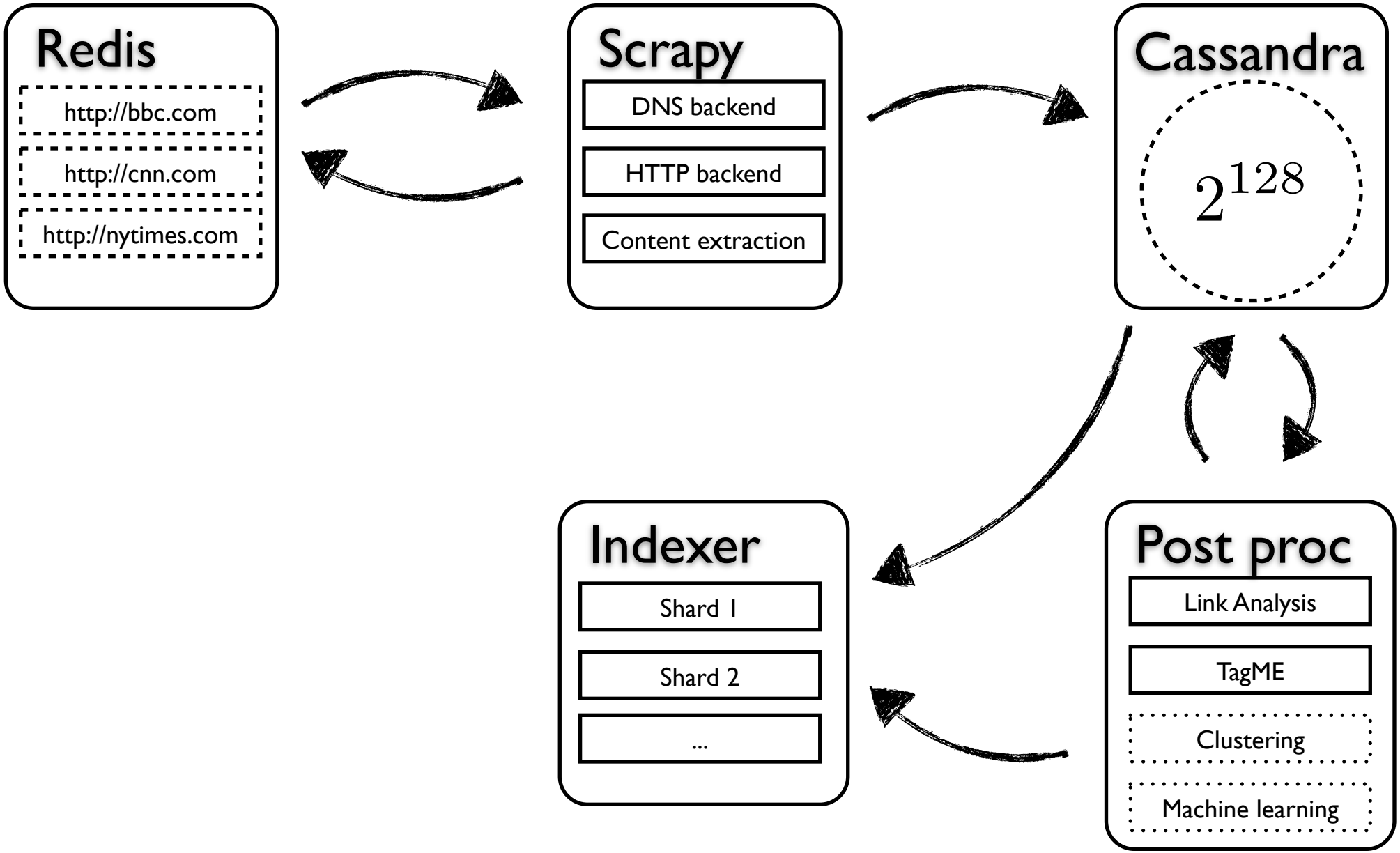
Backend


Redis
http://bbc.com
http://cnn.com
http://nytimes.com
Scrapy
DNS backend
HTTP backend
Content extraction
Cassandra
2^{128}
Indexer
Shard 1
Shard 2
...
Post proc
Link Analysis
TagME
Clustering
Machine learning

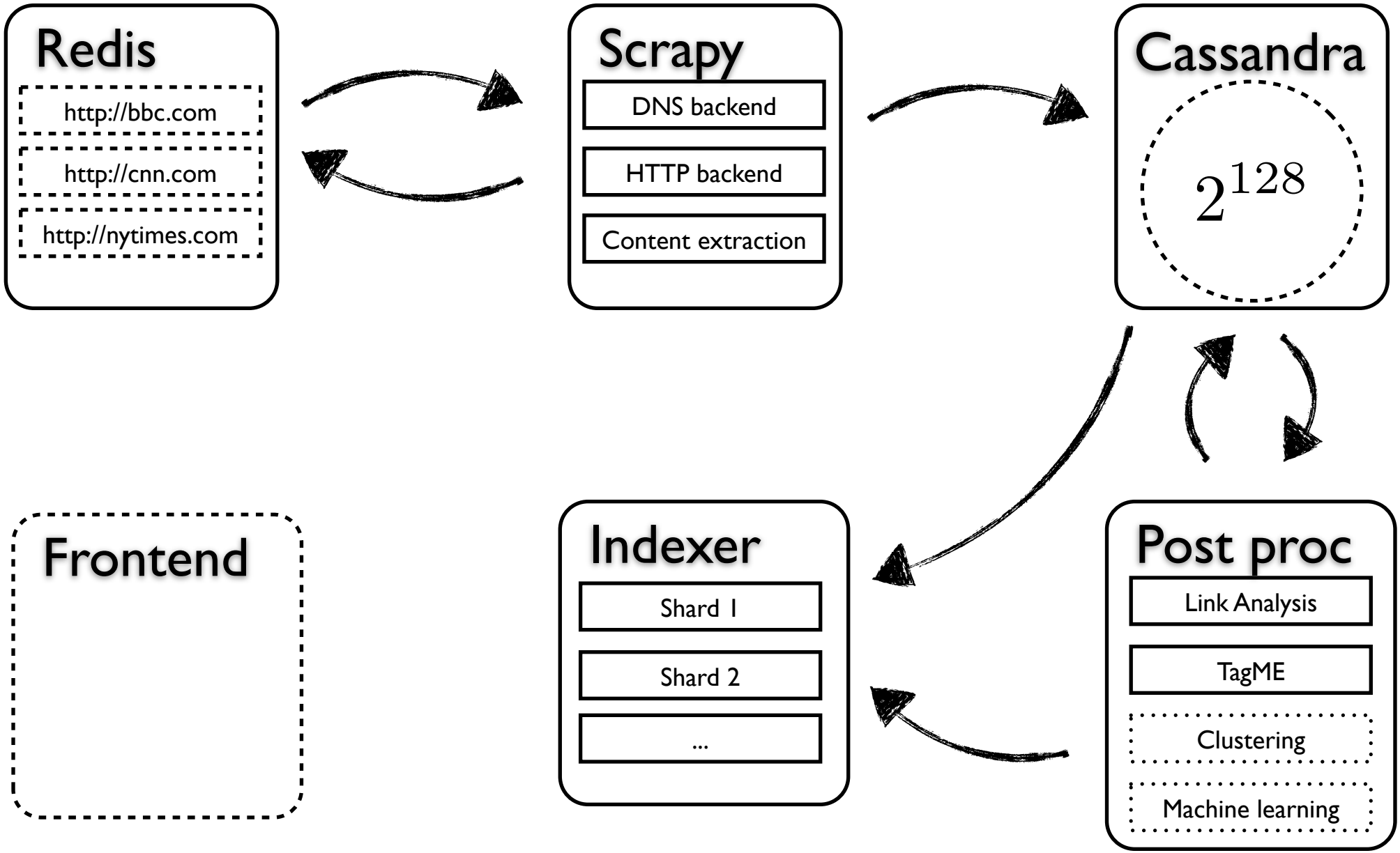
Backend


Redis
http://bbc.com
http://cnn.com
http://nytimes.com
Scrapy
DNS backend
HTTP backend
Content extraction
Cassandra
2^{128}
Indexer
Shard 1
Shard 2
...
Post proc
Link Analysis
TagME
Clustering
Machine learning

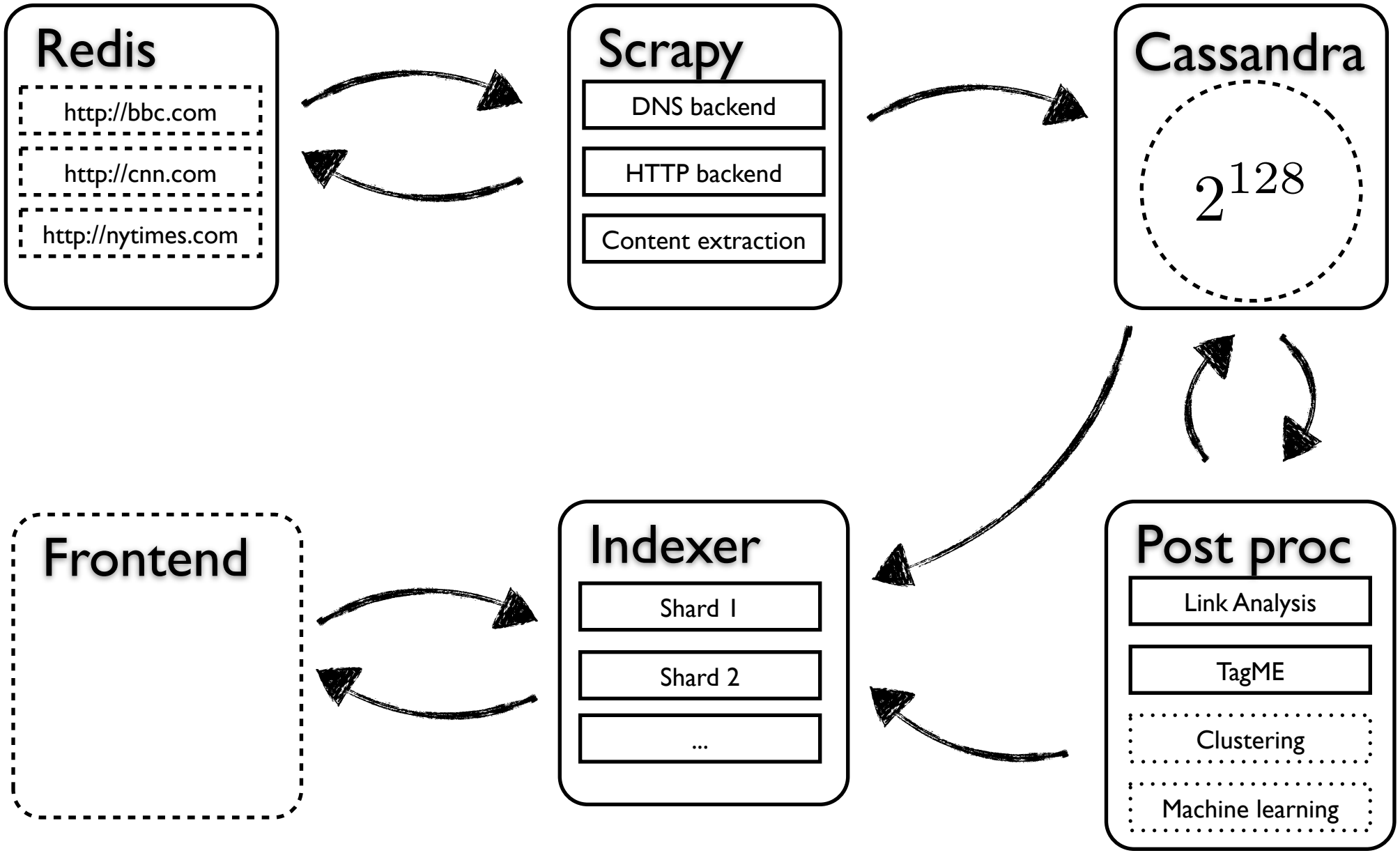
Backend


Redis
http://bbc.com
http://cnn.com
http://nytimes.com
Scrapy
DNS backend
HTTP backend
Content extraction
Cassandra
2^{128}
Indexer
Shard 1
Shard 2
...
Post proc
Link Analysis
TagME
Clustering
Machine learning

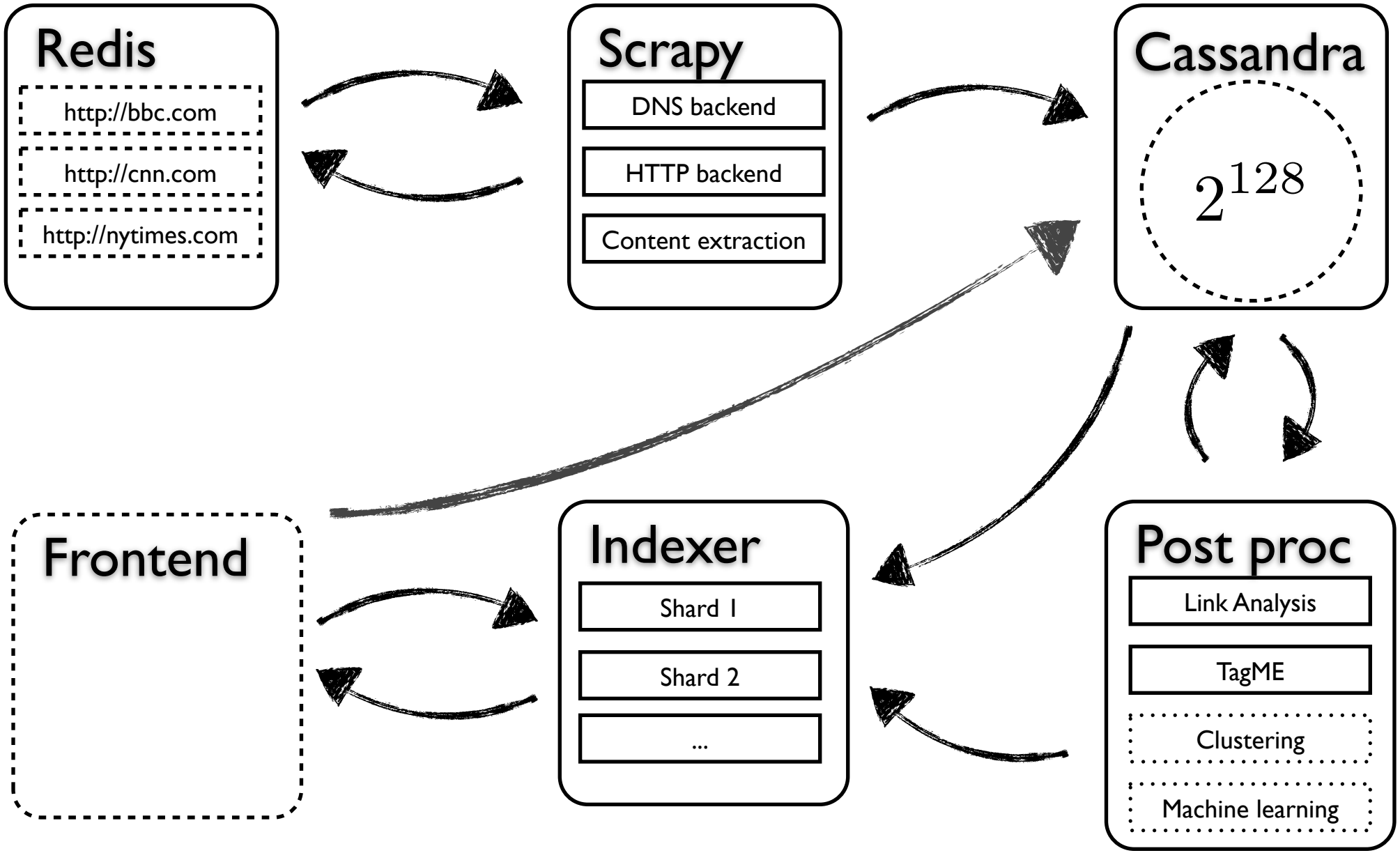
Backend


Redis
http://bbc.com
http://cnn.com
http://nytimes.com
Scrapy
DNS backend
HTTP backend
Content extraction
Cassandra
2^{128}
Frontend
Indexer
Shard 1
Shard 2
...
Post proc
Link Analysis
TagME
Clustering
Machine learning

Backend


Redis
http://bbc.com
http://cnn.com
http://nytimes.com
Scrapy
DNS backend
HTTP backend
Content extraction
Cassandra
2^{128}
Frontend
Indexer
Shard 1
Shard 2
...
Post proc
Link Analysis
TagME
Clustering
Machine learning

Backend


Redis
http://bbc.com
http://cnn.com
http://nytimes.com
Scrapy
DNS backend
HTTP backend
Content extraction
Cassandra
2^{128}
Frontend
Indexer
Shard 1
Shard 2
...
Post proc
Link Analysis
TagME
Clustering
Machine learning

Frontend

- The interface that lets the user interact with the backend
- How to handle million users?
  - Load balancing (DNS, HTTP)
  - Caching (DNS, HTTP)
  - Multi datacenter in different locations

Frontend

- The interface that lets the user interact with the backend
- How to handle million users?
  - Load balancing (DNS, HTTP)
  - Caching (DNS, HTTP)
  - Multi datacenter in different locations

```
nopper@octocat ~ » dig +noall +answer www.google.it
www.google.it.          300     IN      A       173.194.35.23
www.google.it.          300     IN      A       173.194.35.24
www.google.it.          300     IN      A       173.194.35.31
nopper@octocat ~ » dig +noall +answer www.google.it
www.google.it.          258     IN      A       173.194.35.63
www.google.it.          258     IN      A       173.194.35.55
www.google.it.          258     IN      A       173.194.35.56
```

Frontend

- The interface that lets the user interact with the backend
- How to handle million users?
  - Load balancing (DNS, HTTP)
  - Caching (DNS, HTTP)
  - Multi datacenter in different locations

```
nopper@octocat ~ » dig +noall +answer www.google.it
www.google.it.          300     IN      A       173.194.35.23
www.google.it.          300     IN      A       173.194.35.24
www.google.it.          300     IN      A       173.194.35.31
nopper@octocat ~ » dig +noall +answer www.google.it
www.google.it.          258     IN      A       173.194.35.63
www.google.it.          258     IN      A       173.194.35.55
www.google.it.          258     IN      A       173.194.35.56
```

```
➜  ~  ping 8.8.8.8
PING 8.8.8.8 (8.8.8.8) 56(84) bytes of data.
64 bytes from 8.8.8.8: icmp_seq=1 ttl=49 time=16.8 ms
```

Frontend

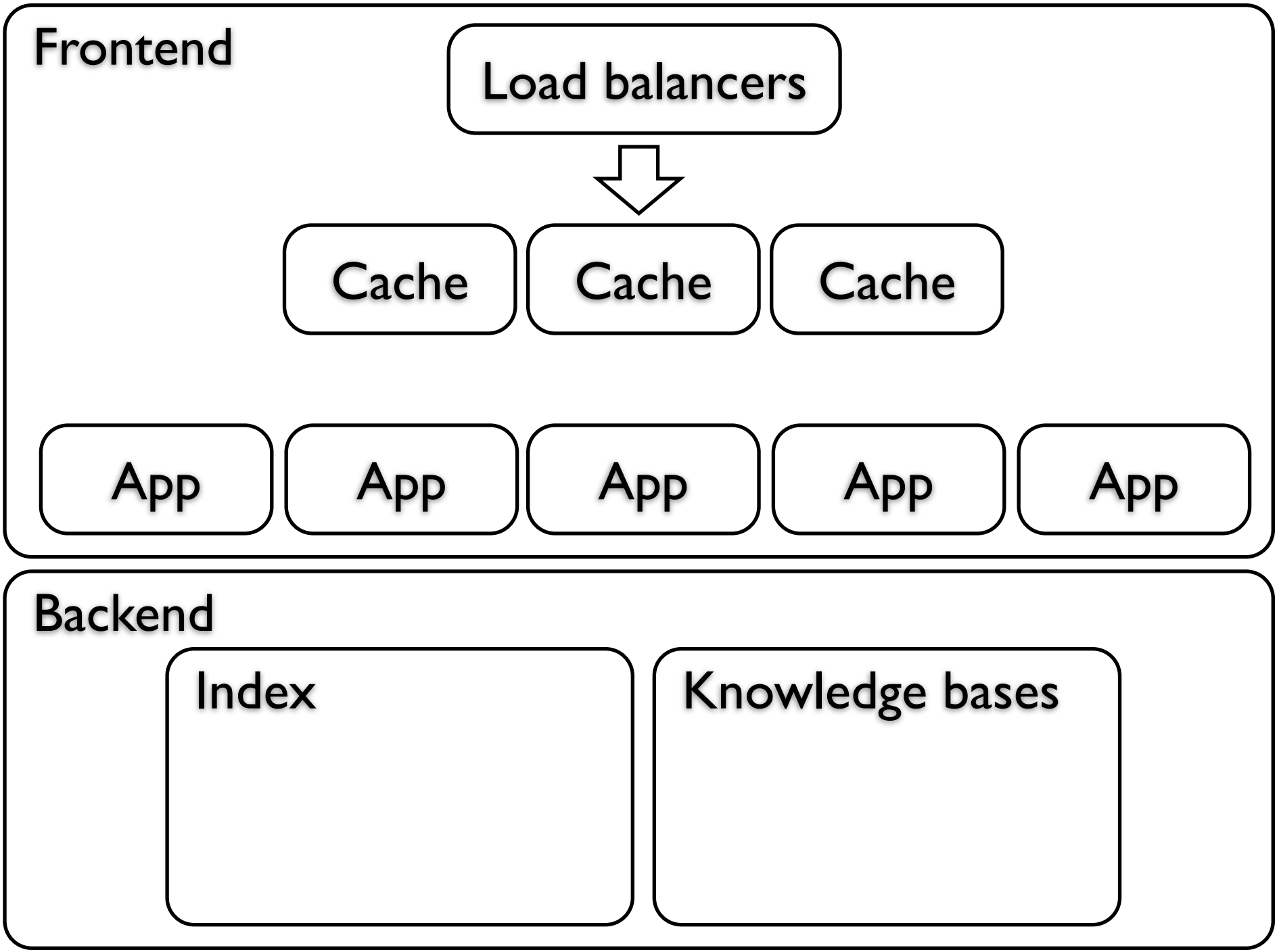
- The interface that lets the user interact with the backend
- How to handle million users?
  - Load balancing (DNS, HTTP)
  - Caching (DNS, HTTP)
  - Multi datacenter in different locations

```
nopper@octocat ~ » dig +noall +answer www.google.it
www.google.it.          300     IN      A       173.194.35.23
www.google.it.          300     IN      A       173.194.35.24
www.google.it.          300     IN      A       173.194.35.31
nopper@octocat ~ » dig +noall +answer www.google.it
www.google.it.          258     IN      A       173.194.35.63
www.google.it.          258     IN      A       173.194.35.55
www.google.it.          258     IN      A       173.194.35.56
```

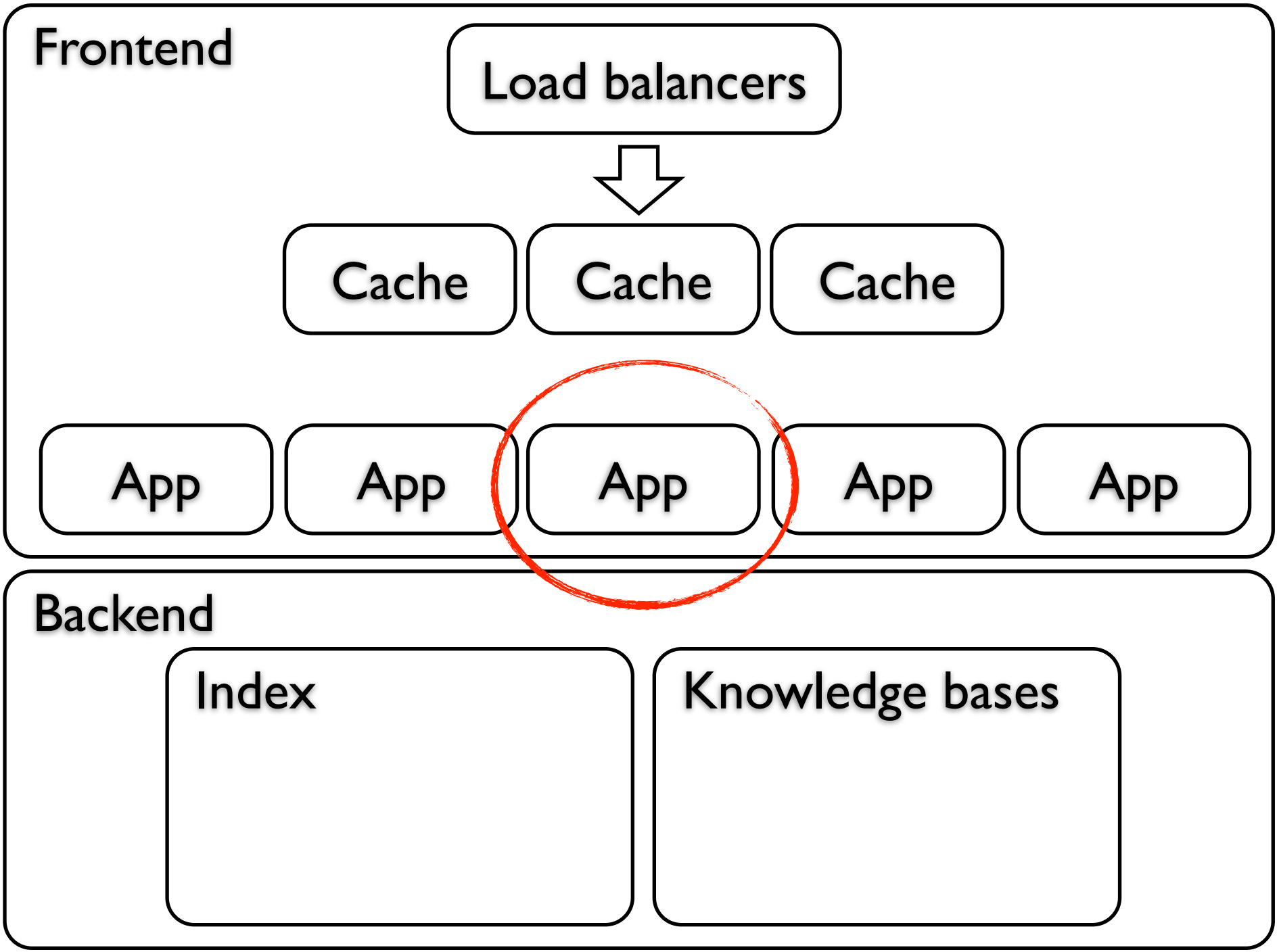
```
➜  ~  ping 8.8.8.8
PING 8.8.8.8 (8.8.8.8) 56(84) bytes of data.
64 bytes from 8.8.8.8: icmp_seq=1 ttl=49 time=16.8 ms
```

```
>>> ((300000 * 0.017) / 2) * (1 - 0.35)
1657.5
```

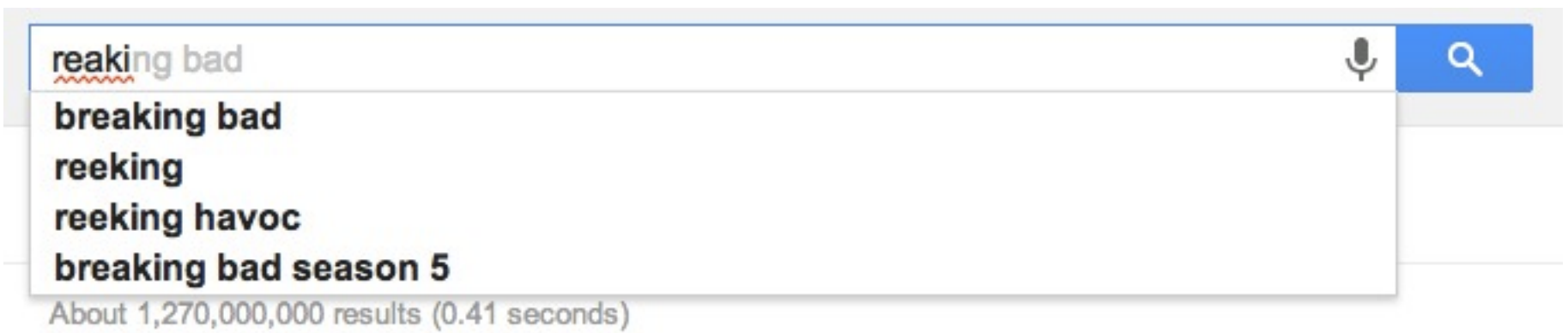
Frontend: high-level



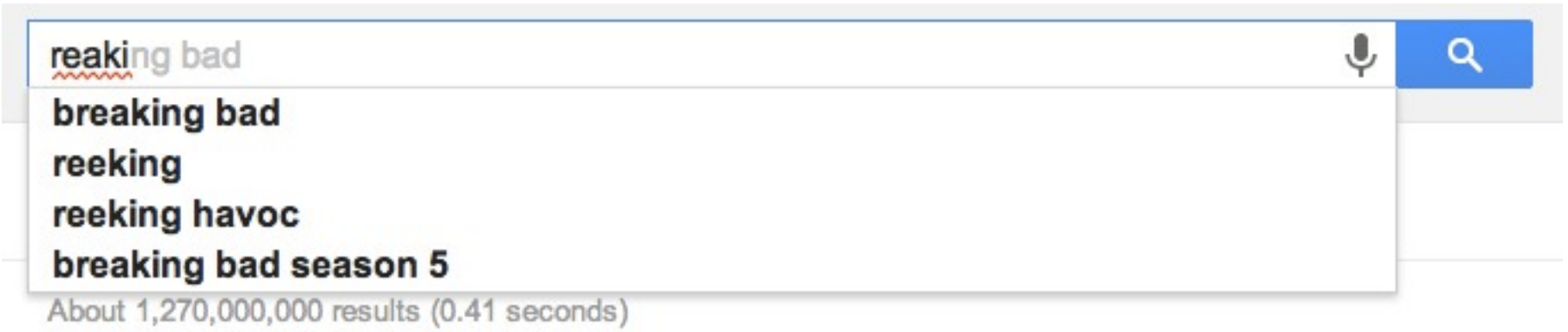
Frontend: high-level



User experience



User experience



User experience

reaking bad

breaking bad
reeking
reeking havoc
breaking bad season 5

About 1,270,000,000 results (0.41 seconds)

Breaking Bad

Television Series

★★★★★ 9.5/10 - IMDb

Breaking Bad is an American crime drama television series created and produced by Vince Gilligan. Set and produced in Albuquerque, New Mexico, Breaking Bad is the story of Walter White, a struggling ... Wikipedia

First episode: January 20, 2008

Final episode: September 29, 2013

Network: AMC

Theme song: Breaking Bad Theme

Writers: Vince Gilligan, Thomas Schnauz, Peter Gould, More

Cast


Bryan Cranston


Aaron Paul
Jesse Pinkman


RJ Mitte
Walter White Jr.


Anna Gunn
Skyler White

User experience

User experience

Scala

Scala

Programming language

Scala is an object-functional programming and scripting language for general software applications, statically typed, designed to concisely express solutions in an elegant, type-safe and lightweight manner. Wikipedia

Feedback / More info

See results about

Scala
Club
Scala is a nightclub in London, England, near King's Cross railway ...

SCALA

Scala, Inc
Software company
Scala, Inc is a producer of multimedia software. Founded in 1987, Scala is ...

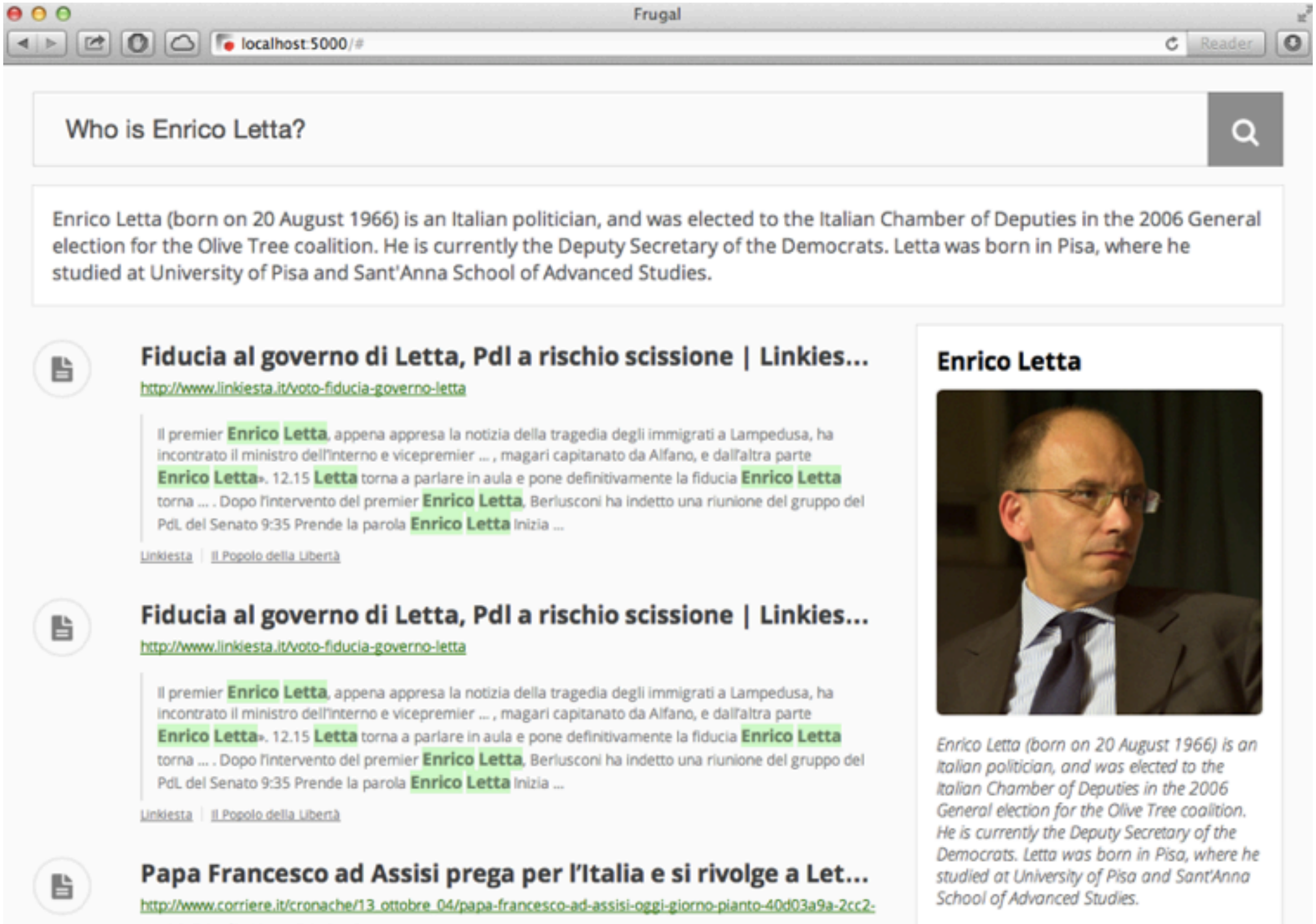
Scala & Kolacny Brothers
La Scala & Kolacny Brothers is a Belgian women's choir, conducted by ...

Teatro alla Scala
Theatre in Milan, Italy
La Scala is a world-renowned opera house in Milan, Italy. The theatre was ...

Frontend

- A simple web application
  - Semantic-UI for the base CSS
  - jQuery to make the application dynamic (AJAX)
  - Typeahead.js
- It works by forwarding JSON request to a frontend server implemented in Python
- The frontend interacts with Elasticsearch instances, DBPedia and TagME

Results


Frugal
localhost:5000/#
Reader
Who is Enrico Letta?
Enrico Letta (born on 20 August 1966) is an Italian politician, and was elected to the Italian Chamber of Deputies in the 2006 General election for the Olive Tree coalition. He is currently the Deputy Secretary of the Democrats. Letta was born in Pisa, where he studied at University of Pisa and Sant'Anna School of Advanced Studies.
Fiducia al governo di Letta, Pdl a rischio scissione | Linkies...
http://www.linkiesta.it/voto-fiducia-governo-letta
Il premier Enrico Letta, appena appresa la notizia della tragedia degli immigrati a Lampedusa, ha incontrato il ministro dell'Interno e vicepremier ... , magari capitanato da Alfano, e dall'altra parte Enrico Letta». 12.15 Letta torna a parlare in aula e pone definitivamente la fiducia Enrico Letta torna ... . Dopo l'intervento del premier Enrico Letta, Berlusconi ha indetto una riunione del gruppo del PdL del Senato 9:35 Prende la parola Enrico Letta Inizia ...
Linkiesta | Il Popolo della Libertà
Fiducia al governo di Letta, Pdl a rischio scissione | Linkies...
http://www.linkiesta.it/voto-fiducia-governo-letta
Il premier Enrico Letta, appena appresa la notizia della tragedia degli immigrati a Lampedusa, ha incontrato il ministro dell'Interno e vicepremier ... , magari capitanato da Alfano, e dall'altra parte Enrico Letta». 12.15 Letta torna a parlare in aula e pone definitivamente la fiducia Enrico Letta torna ... . Dopo l'intervento del premier Enrico Letta, Berlusconi ha indetto una riunione del gruppo del PdL del Senato 9:35 Prende la parola Enrico Letta Inizia ...
Linkiesta | Il Popolo della Libertà
Papa Francesco ad Assisi prega per l'Italia e si rivolge a Let...
http://www.corriere.it/cronache/13_ottobre_04/papa-francesco-ad-assisi-oggi-giorno-pianto-40d03a9a-2cc2-
Enrico Letta
Enrico Letta (born on 20 August 1966) is an Italian politician, and was elected to the Italian Chamber of Deputies in the 2006 General election for the Olive Tree coalition. He is currently the Deputy Secretary of the Democrats. Letta was born in Pisa, where he studied at University of Pisa and Sant'Anna School of Advanced Studies.

References

- Scrapy
  - http://scrapy.org/
- Cassandra
  - http://cassandra.apache.org/
- Spark
  - http://spark.incubator.apache.org/
- Elasticsearch
  - http://www.elasticsearch.org/
- TagME
  - http://acube.di.unipi.it/tagme