Buscas poderosas em bilhões de documentos?

Seu sistema pode prover isso de forma escalável e resiliente com o **Elasticsearch**

Matheus de Faria Moraes

Agenda

- What is Elasticsearch;
- Use Cases;
- Basic Concepts;
- Document and Index;
- Cluster and Nodes;
- Primary Shards and Replica Shards;
- Near Real Time (NRT);
- Demo.

whoami

I am **Matheus Moraes**

Software Engineer **@Sensedia**

Java, Golang, NoSQL and Microservices enthusiast

What is Elasticsearch?

What is Elasticsearch?

- Full-text search and analytics engine;
- Highly scalable;
- Open-source;
- Store, search, and analyze big volumes of data in **near real time**;
- REST APIs;
- Good documentation;
- Apache Lucene.

Use Cases

GitHub	Facebook	Uber	Netflix	SoundCloud	The New York TImes	Microsoft 2015
Accelerating software development	Delivering a better help experience for over a billion users	Aggregating business metrics to control critical marketplace behaviors	Ensuring message delivery and operational excellence	Helping users find the sounds that move them	Providing search for all 164 years of published articles	Providing search on Azure and powering Social Dynamics

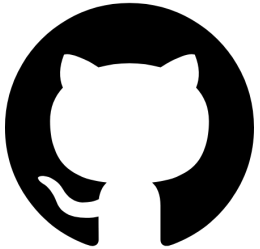
Deezer	Globo.com	MercadoLibre	Tinder	Vimeo
Connecting 12 million listeners to the music they want to hear	Delivering the news, entertainment, and sports content Brazilians want	Powering Latin America's premier online marketplace	Connecting people around the world	Powering the search for the video you want

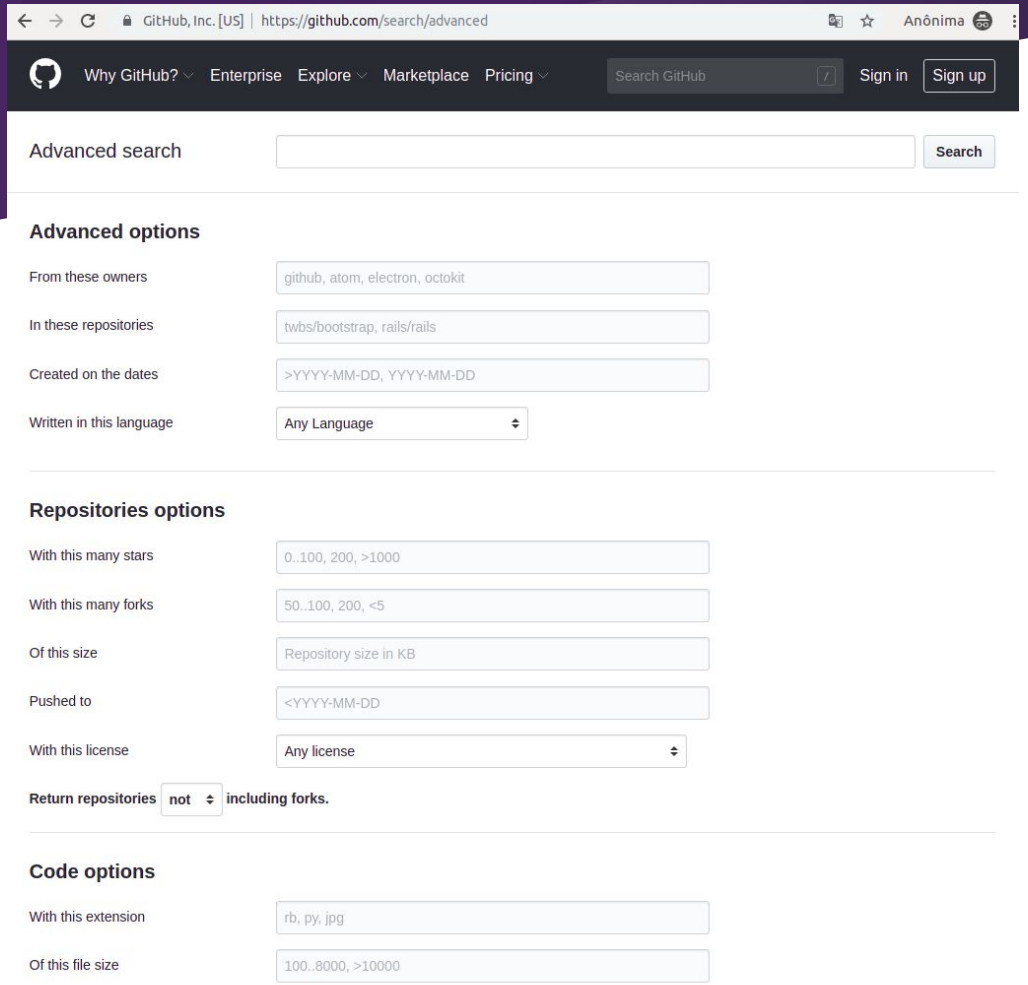
2 **billion** documents

8 million code repositories

4 million active users

~ 300 search / minute



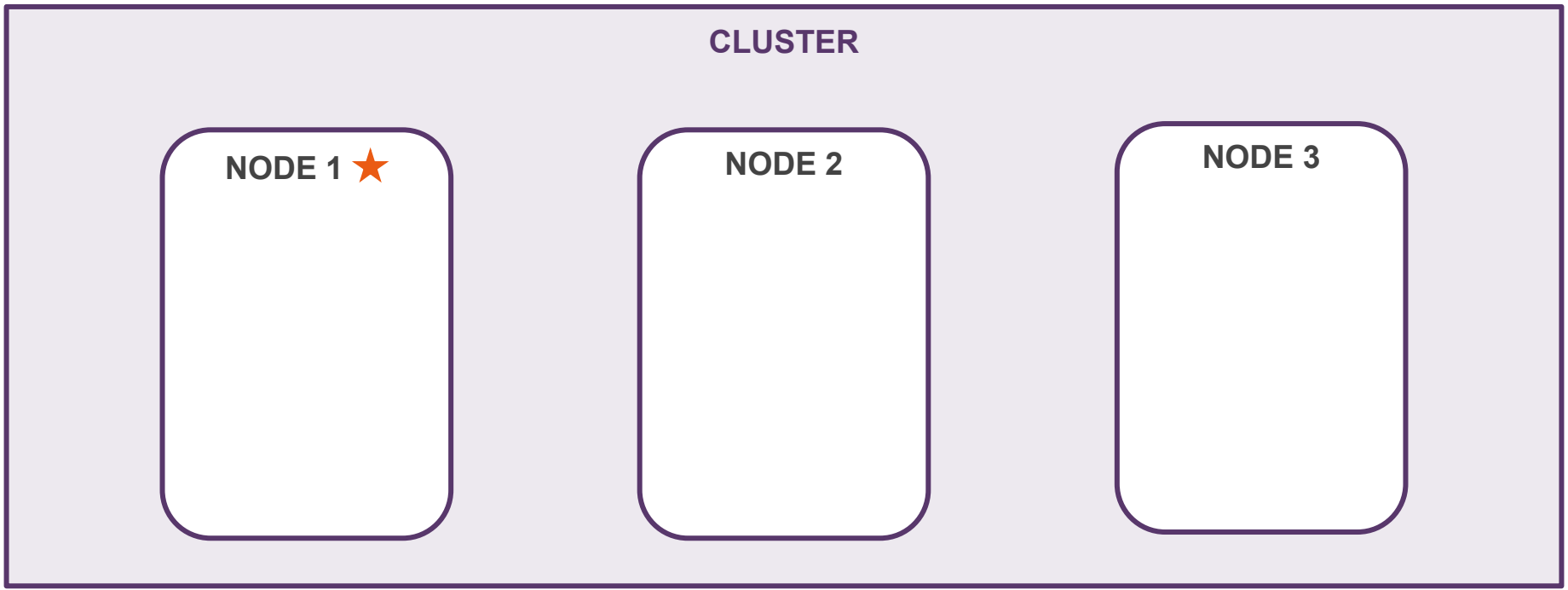

GitHub, Inc. [US] | https://github.com/search/advanced
Anônima
Why GitHub?
Enterprise
Explore
Marketplace
Pricing
Search GitHub
/
Sign in
Sign up
Advanced search
Search
Advanced options
From these owners
github, atom, electron, octokit
In these repositories
twbs/bootstrap, rails/rails
Created on the dates
>YYYY-MM-DD, YYYY-MM-DD
Written in this language
Any Language
Repositories options
With this many stars
0..100, 200, >1000
With this many forks
50..100, 200, <5
Of this size
Repository size in KB
Pushed to
<YYYY-MM-DD
With this license
Any license
Return repositories not including forks.
Code options
With this extension
rb, py, jpg
Of this file size
100..8000, >10000

Basic Concepts

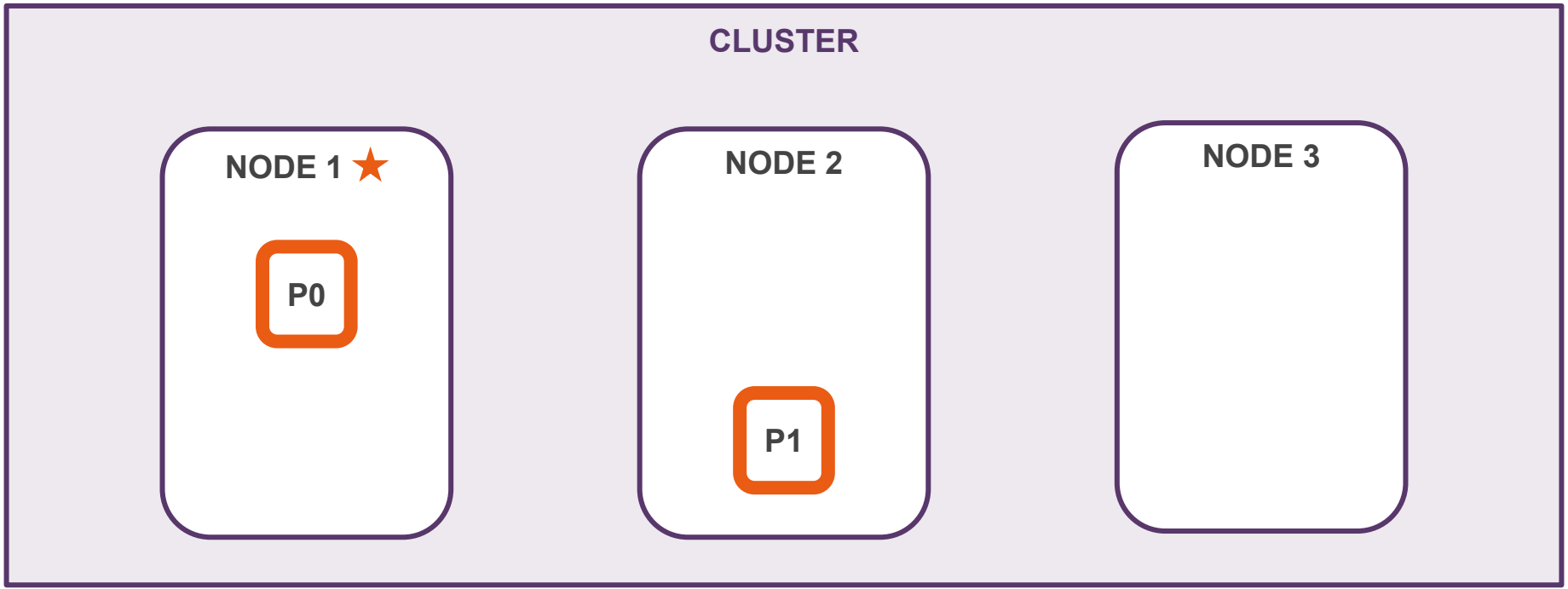
Document and Index

```
curl -X PUT localhost:9200/cities/_doc/1 \
  -H 'Content-Type: application/json' \
  -d '{
          "city": "Tanabi",
          "state": "SP",
          "country": "BR",
          "population": 25000
       }'
```

Cluster and Nodes



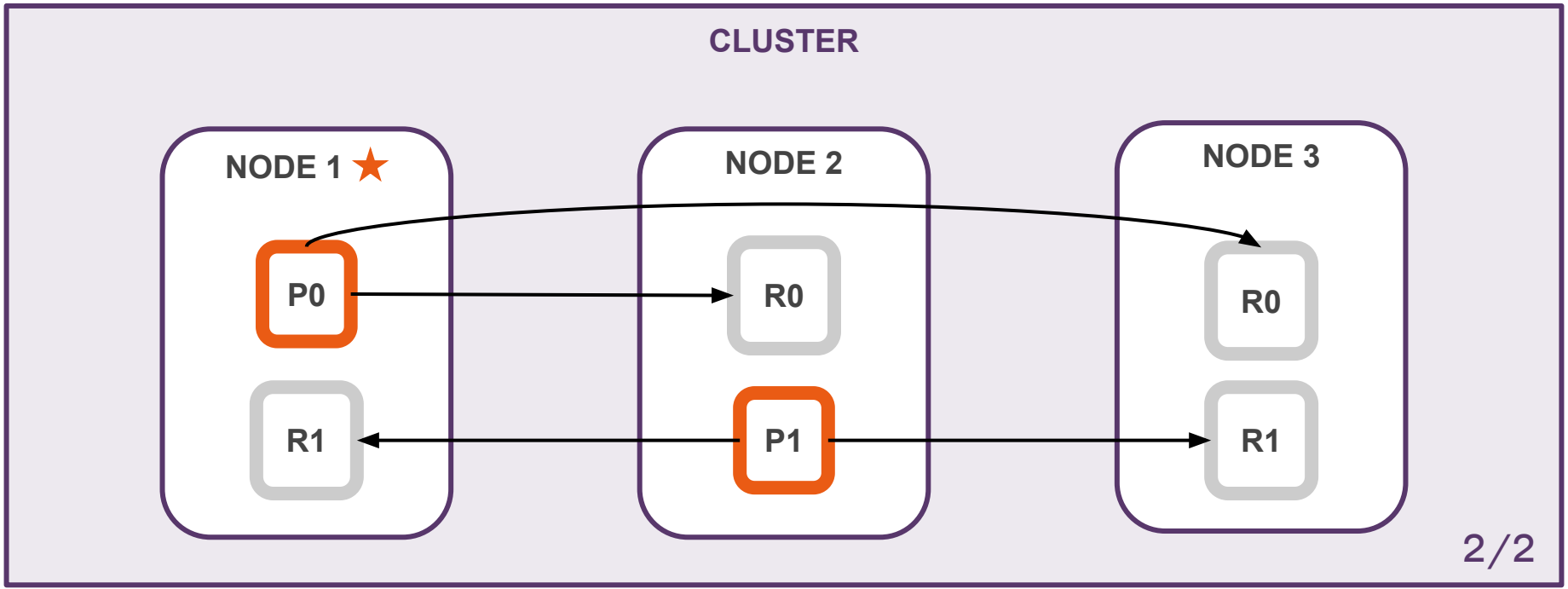
Primary Shards



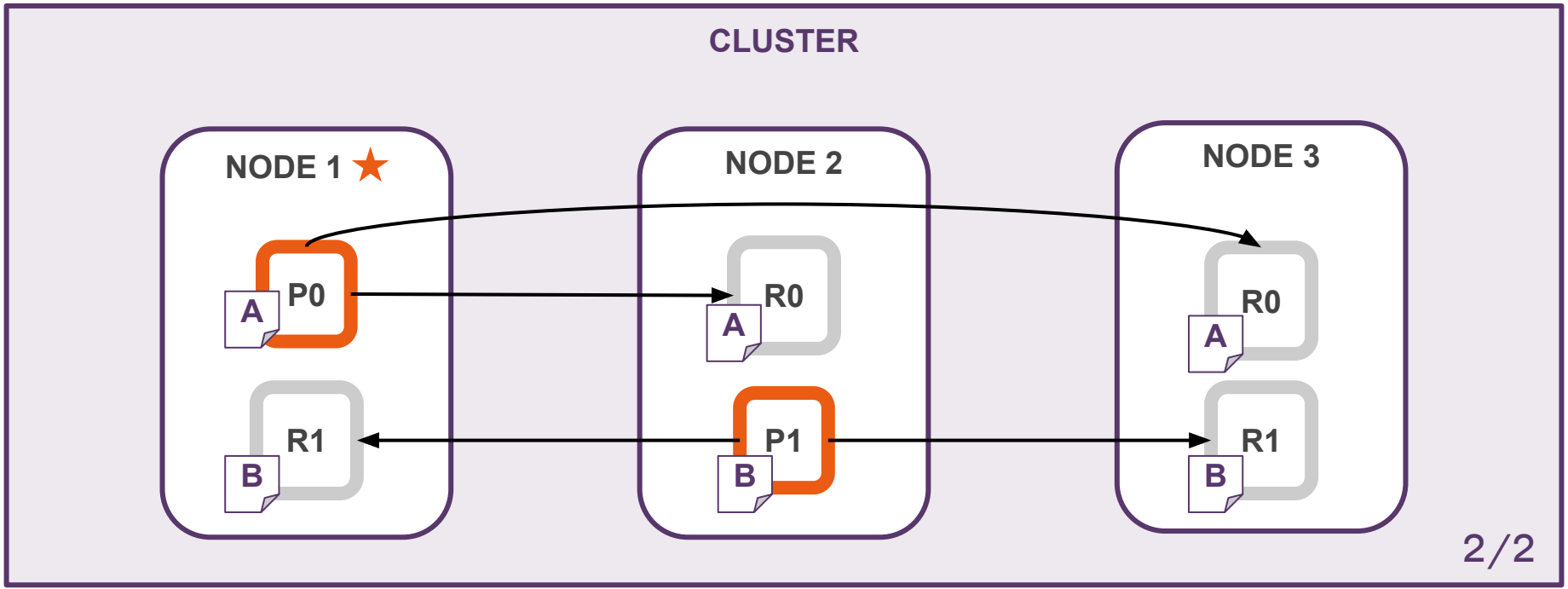
Primary Shard Benefits

- Elasticity
- Horizontal Scaling

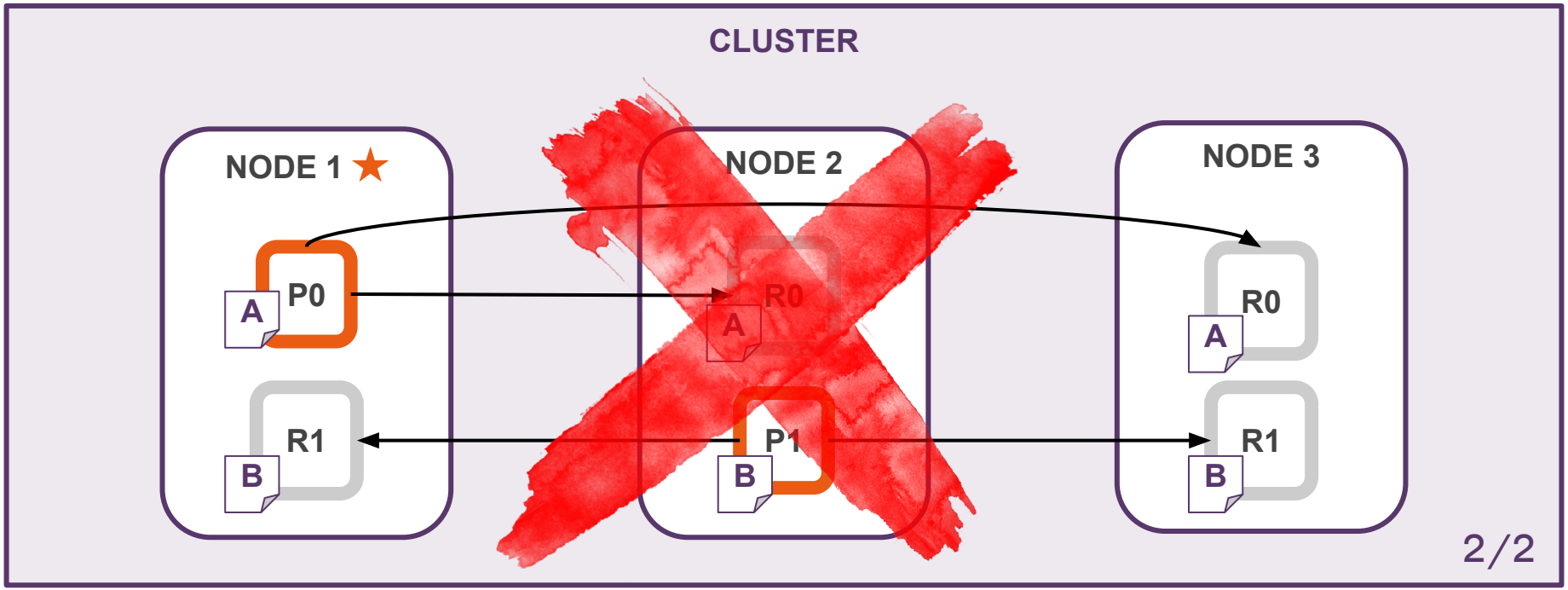
Replica Shards



Cluster, Nodes and Shards



Cluster, Nodes and Shards



Replica Shard Benefits

- H.A.
- Resilience
- Search Throughput

Topologies

- Default 7.0 **1 / 1**
- Old Default **5 / 1**
- Search performance **1 / 10**
- Index performance **20 / 1**

Index creation with shards

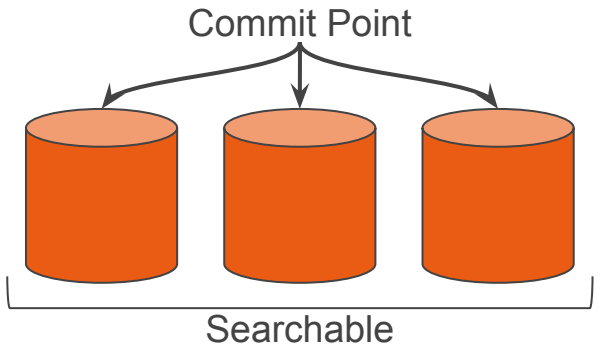
```
curl -X PUT localhost:9200/cities \
  -H 'Content-Type: application/json' \
  -d '{
          "settings": {
              "number_of_shards": 2,
              "number_of_replicas": 1
           }
       }'
```

Searchable and Persistent Documents

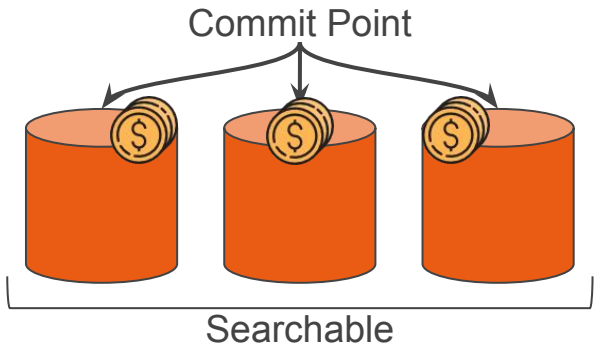
NRT

```
curl -X PUT localhost:9200/cities/_doc/1 \
  -H 'Content-Type: application/json' \
  -d '{
    "city": "Tanabi",
    "state": "SP",
    "country": "BR",
    "population": 25000
}' && \
curl -X GET localhost:9200/cities/_search?pretty&q=name:Tanabi
```

Search by segment (Lucene)



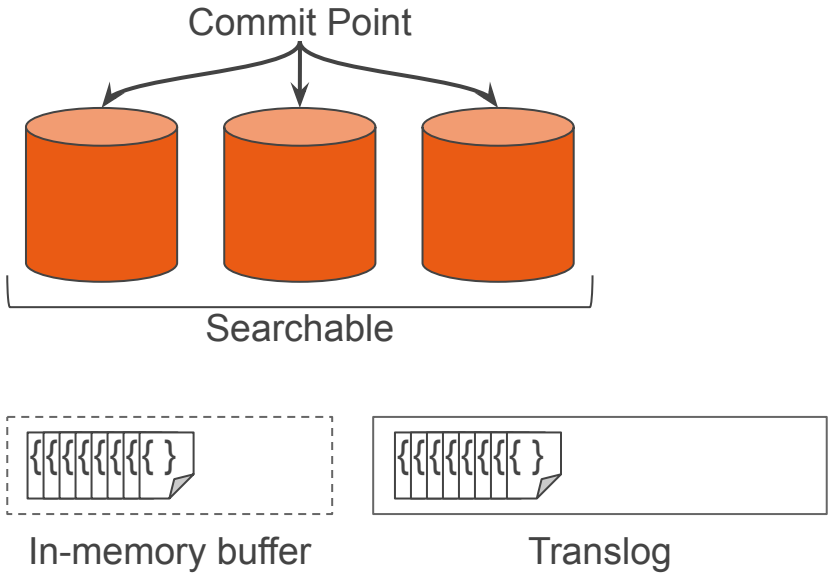
Lucene commits are expensive



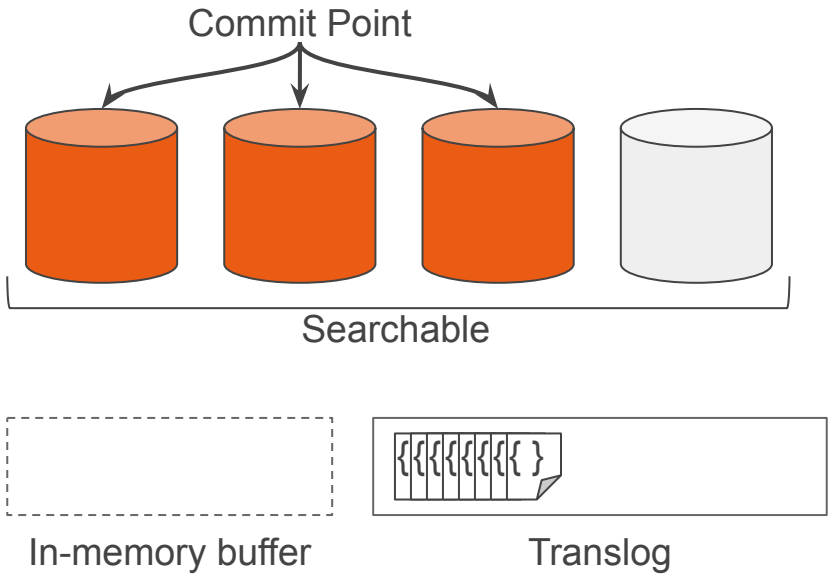
- `fsync`
- Disk

In-memory buffer and Translog

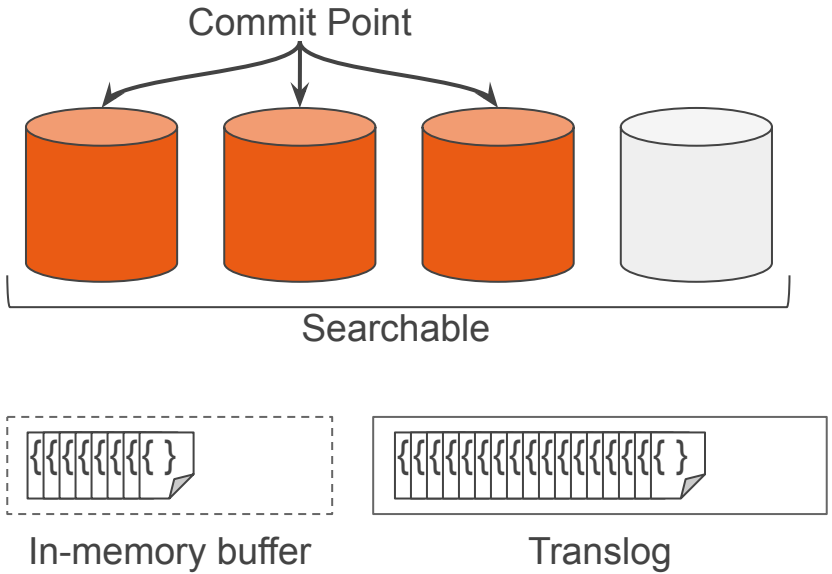
1. Documents are indexed



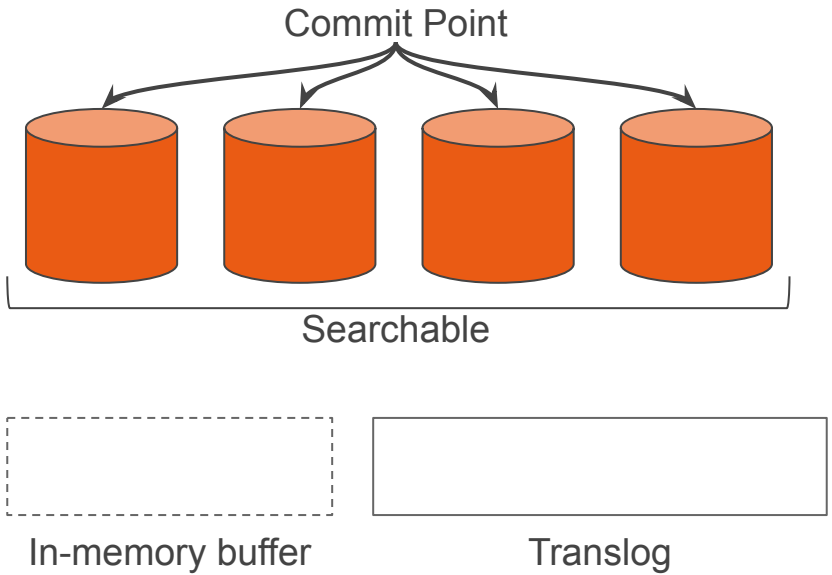
2. Refresh



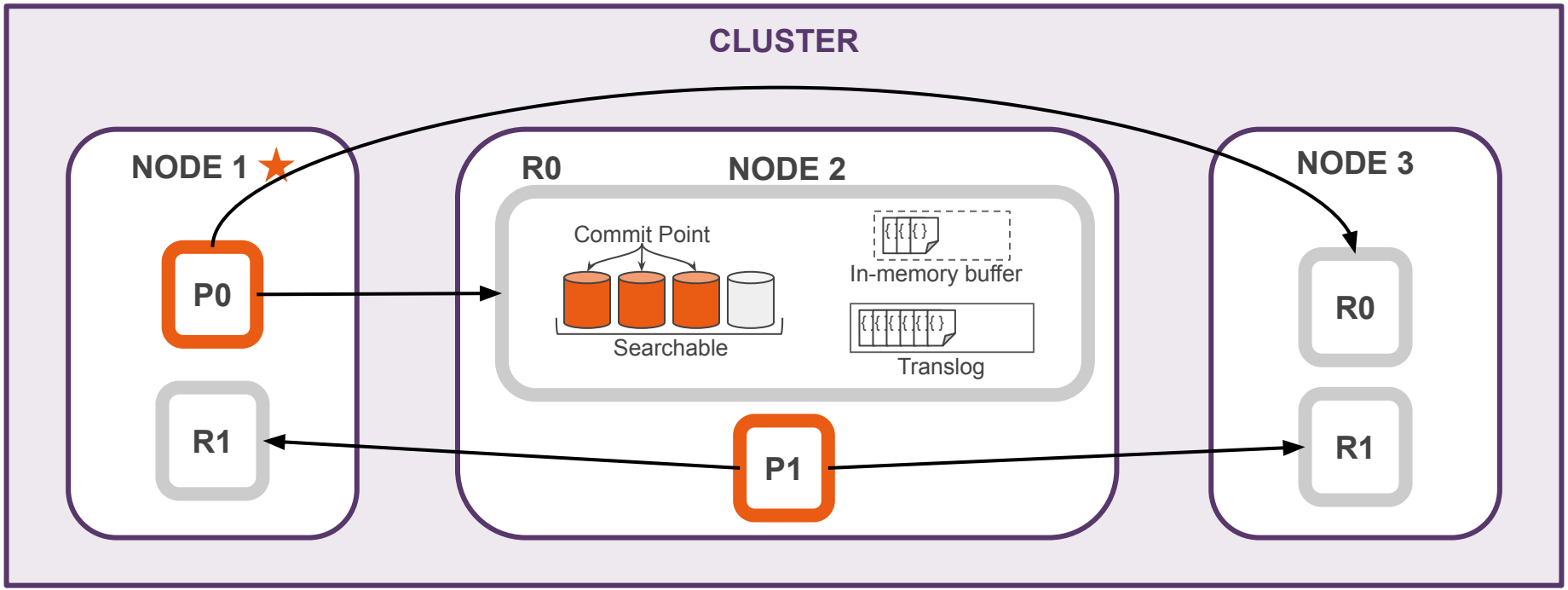
3. The translog keeps accumulating documents



4. Flush (Lucene commit)



Big picture


CLUSTER
NODE 1
P0
R1
R0
NODE 2
Commit Point
Searchable
In-memory buffer
Translog
P1
NODE 3
R0
R1

Refresh interval

```
curl -X PUT localhost:9200/cities/_settings \
  -H 'Content-Type: application/json' \
  -d '{
    "index" : {
       "refresh_interval" : "3s"
    }
}'
```

Refresh

`?refresh` (Index, Update, Delete, and Bulk)

- Empty or `true`
- `wait_for`
- `false` (default)

```
POST cities/_refresh
```

Demo

matheusfm/elasticsearch-demo

Thank you!

 matheusfm

 matheusfm

 mfariam