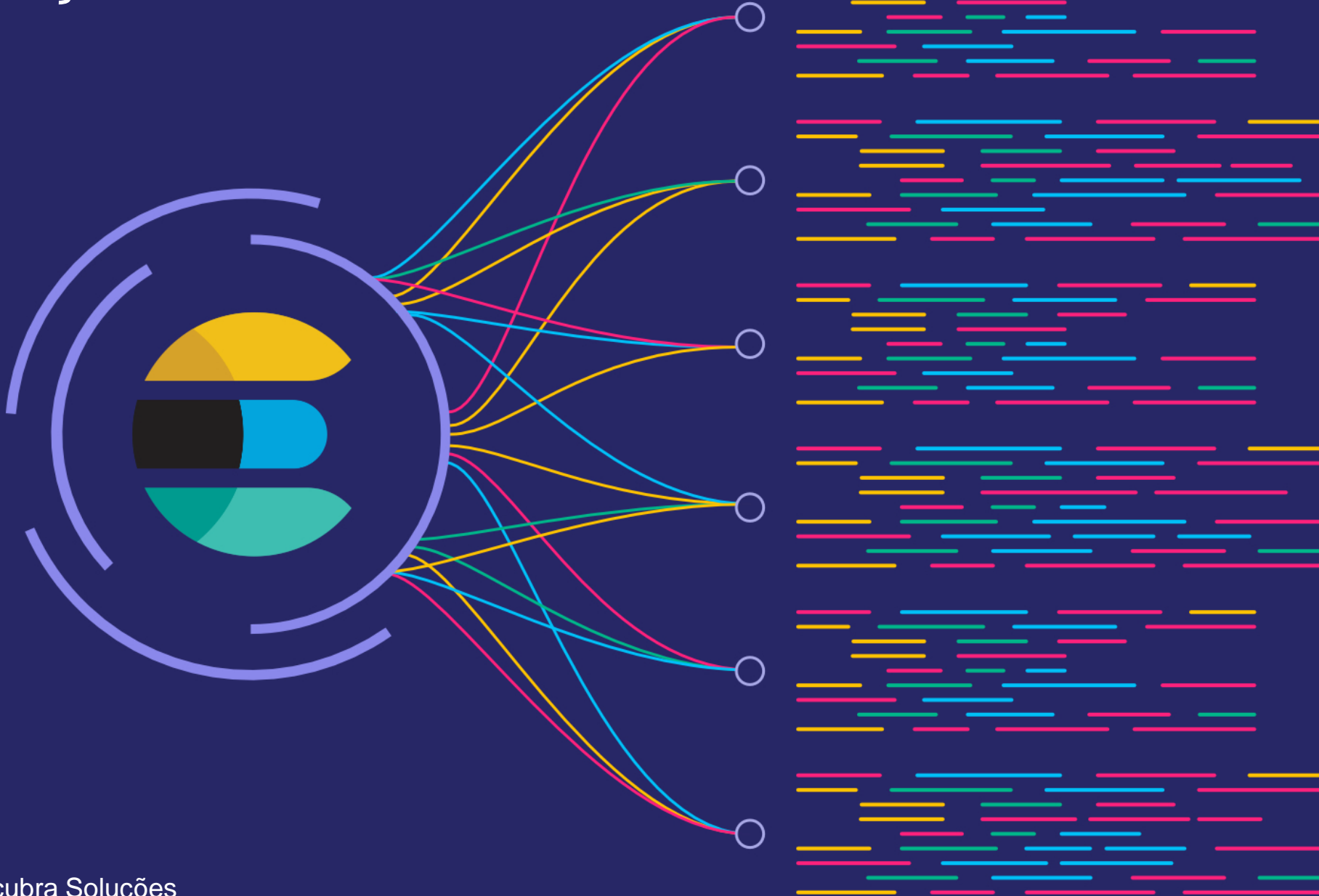


Introdução ao Elasticsearch



O que é o Elasticsearch?

Elastic Stack

Conjunto de ferramentas para coletar, buscar, analisar e visualizar dados de diferentes fontes



elastic stack

Kibana
(Explorar, visualizar)



Elasticsearch
(Armazenar, buscar, analisar)



Integrations
(Conectar, coletar, alertar)



O que é o Elasticsearch?

Elasticsearch
(Armazenar, buscar, analisar)



Elasticsearch é um **mecanismo de busca** em tempo real, **baseado em documentos**, que permite o **armazenamento**, a **pesquisa** e a **análise** de grandes volumes de dados rapidamente.

Mecanismos de buscas

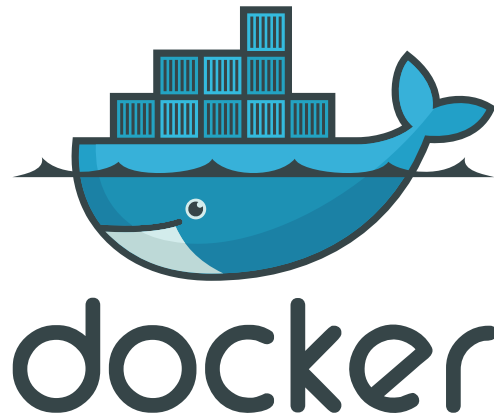
Um mecanismo de busca é uma ferramenta de software projetada para buscar informações de forma rápida e eficiente em um grande conjunto de dados, indexando e organizando essas informações para que possam ser consultadas facilmente.

Boa experiência de busca = Obter resultados relevantes de forma rápida, independente da escala

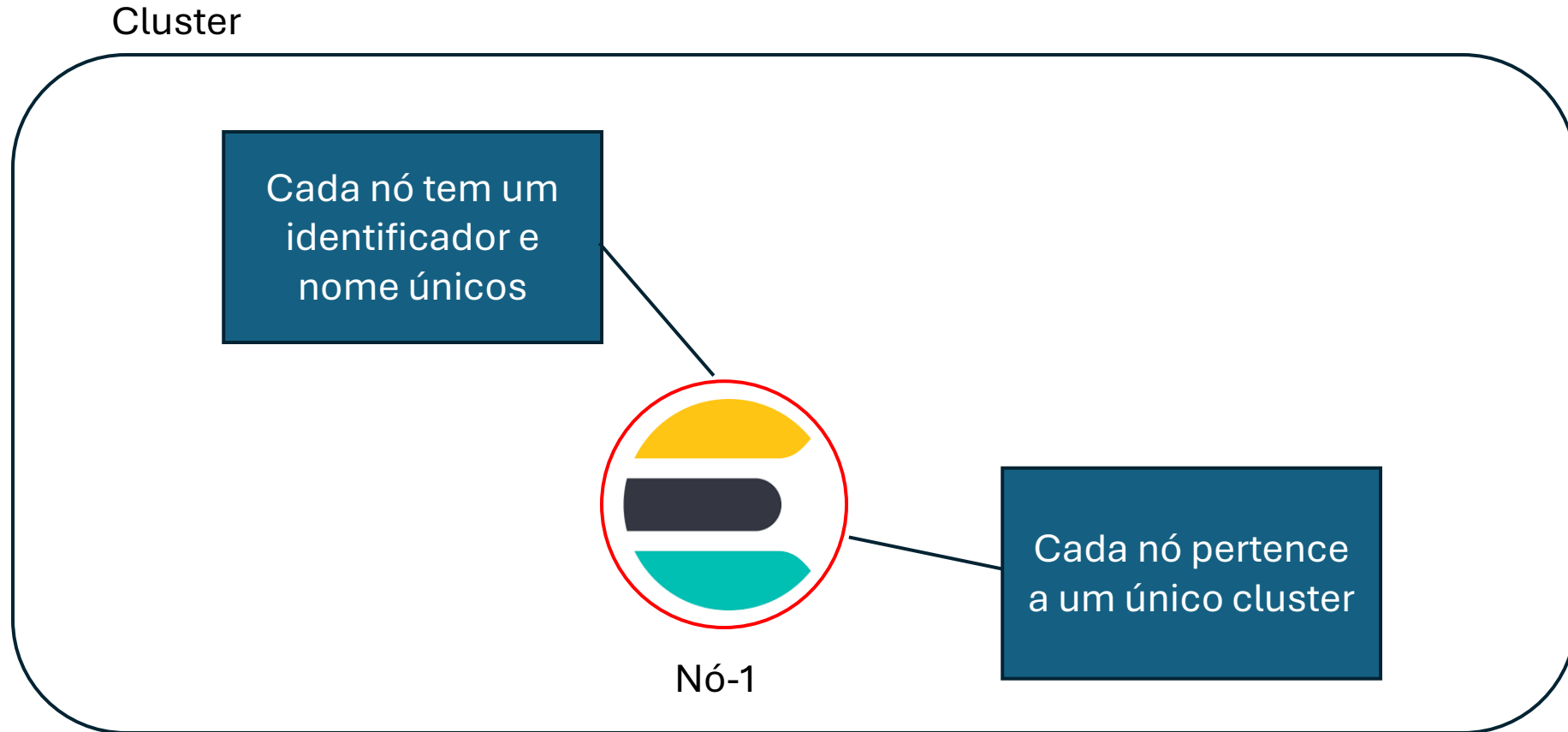


Aplicações que utilizam o Elasticsearch

Uber

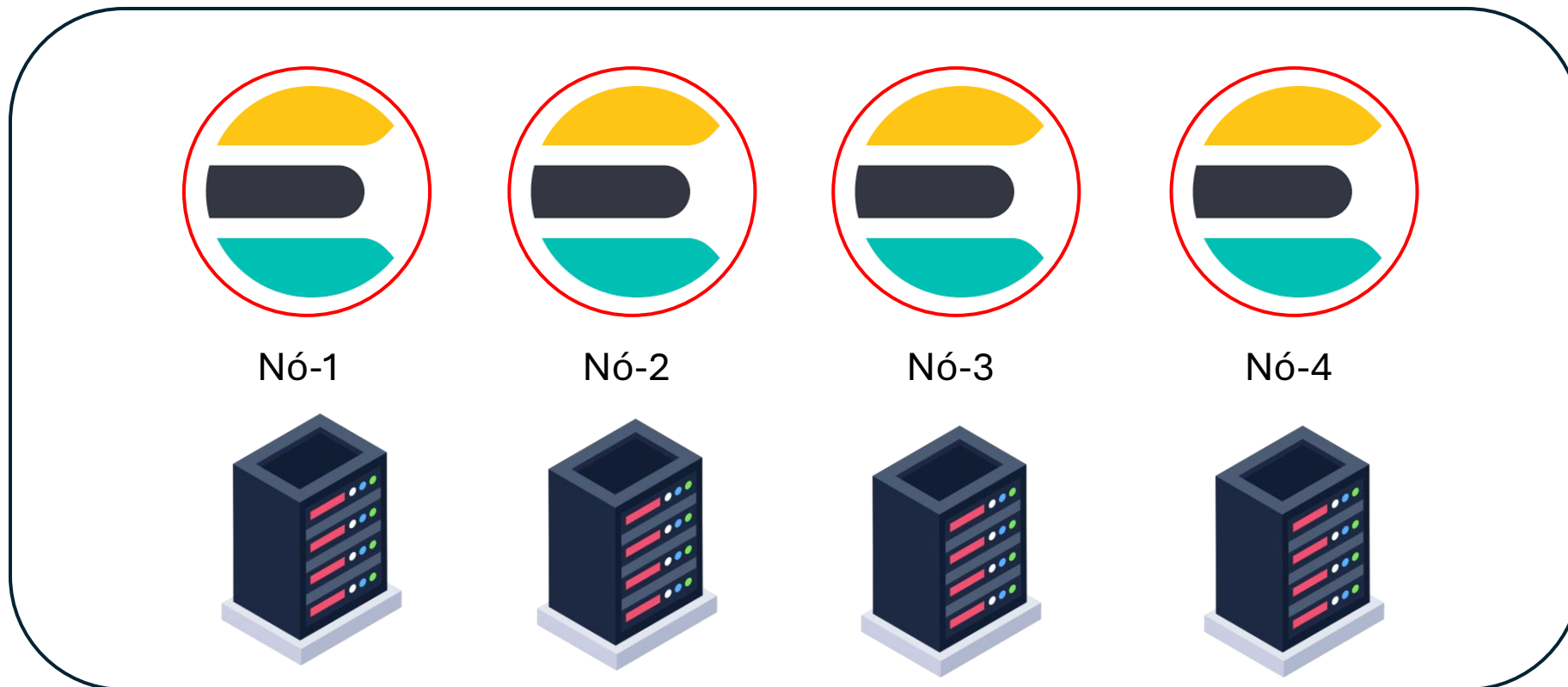


Arquitetura do Elasticsearch



Arquitetura do Elasticsearch

Cluster



Dados são armazenados em documentos

```
{  
  "nome": "Banana Prata",  
  "categoria": "Frutas",  
  "preco": "R$ 6,37"  
}
```

Um documento é um objeto JSON que contém os dados que serão armazenados no elasticsearch

Documentos são agrupados em índices

Índice de hortifruti

```
{  
  "nome": "Banana Prata",  
  "categoria": "Frutas",  
  "preco": "R$ 6,37"  
}
```

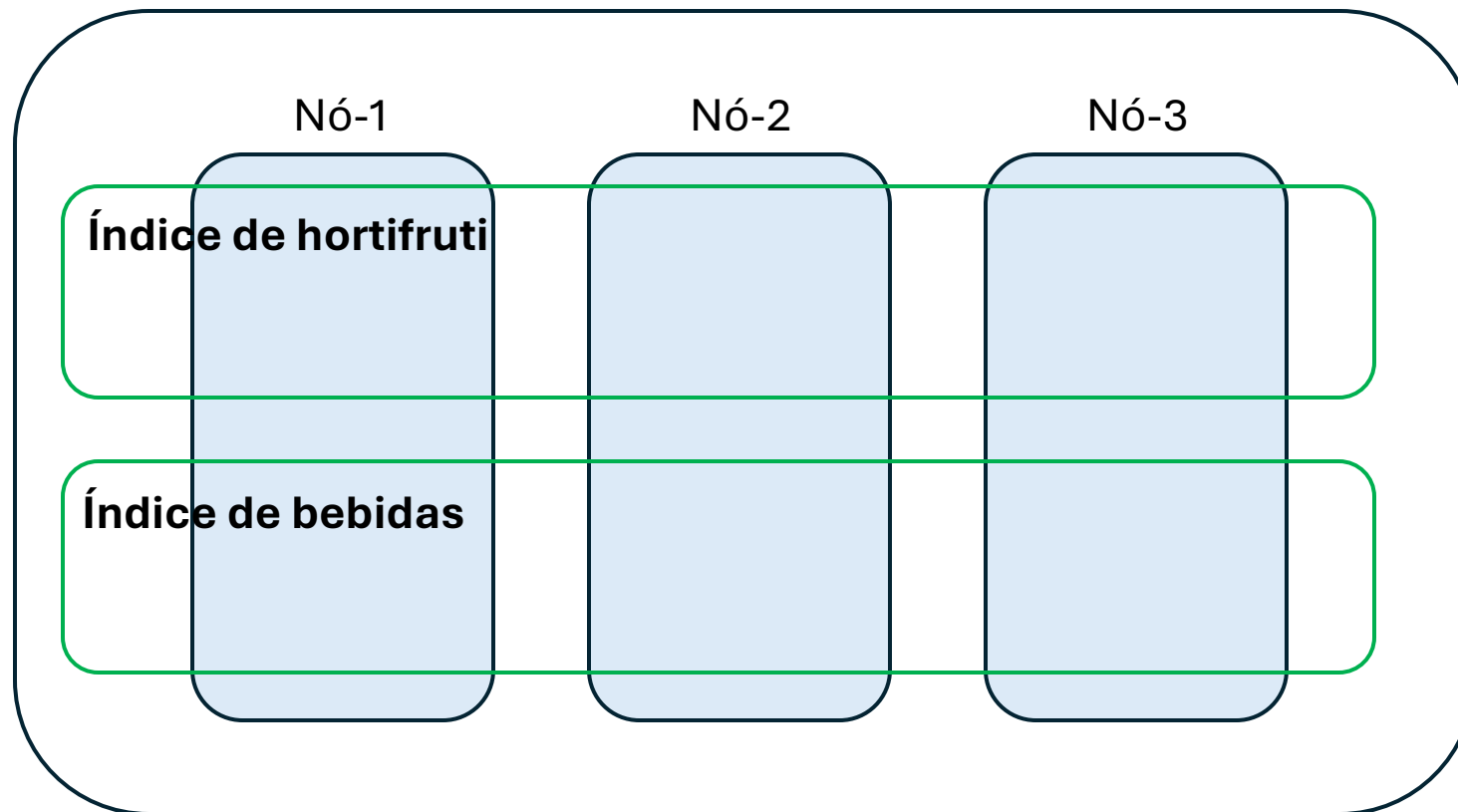
```
{  
  "nome": "Batata inglesa",  
  "categoria": "Legume",  
  "preco": "R$ 7,89"  
}
```

Índice de bebidas

```
{  
  "nome": "Spaten",  
  "categoria": "Cerveja",  
  "preco": "R$ 4,50",  
  "tipo": "Lata"  
}
```

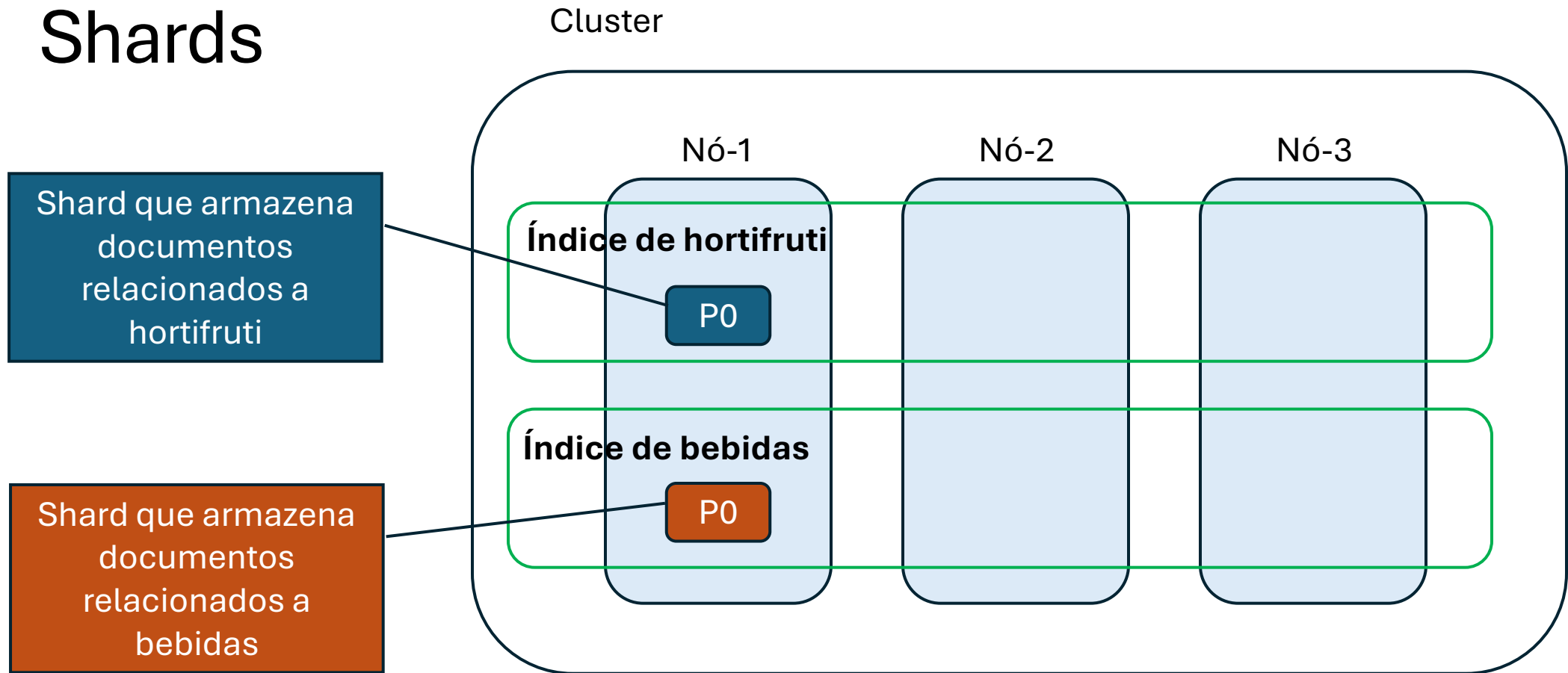
```
{  
  "nome": "Casillero del Diablo - Red Blend",  
  "categoria": "Vinho",  
  "preco": "R$ 49,90",  
  "tipo": "Garrafa"  
}
```

Cluster



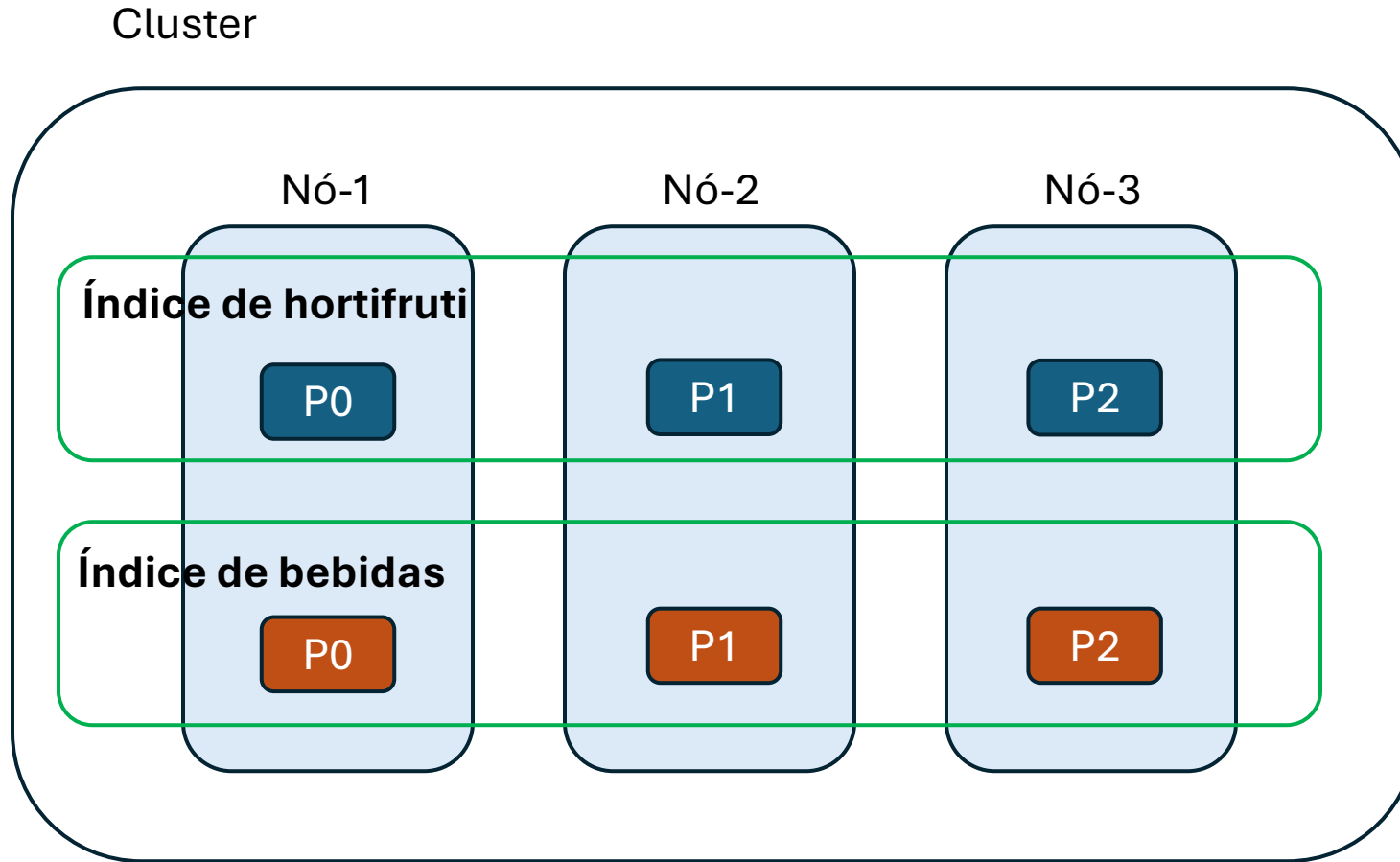
O índice não armazena os documentos. Ele é um recurso virtual que mantém um controle de onde os documentos estão armazenados!

Shards



Shard é onde os dados são armazenados e a busca é realizada.

Shards

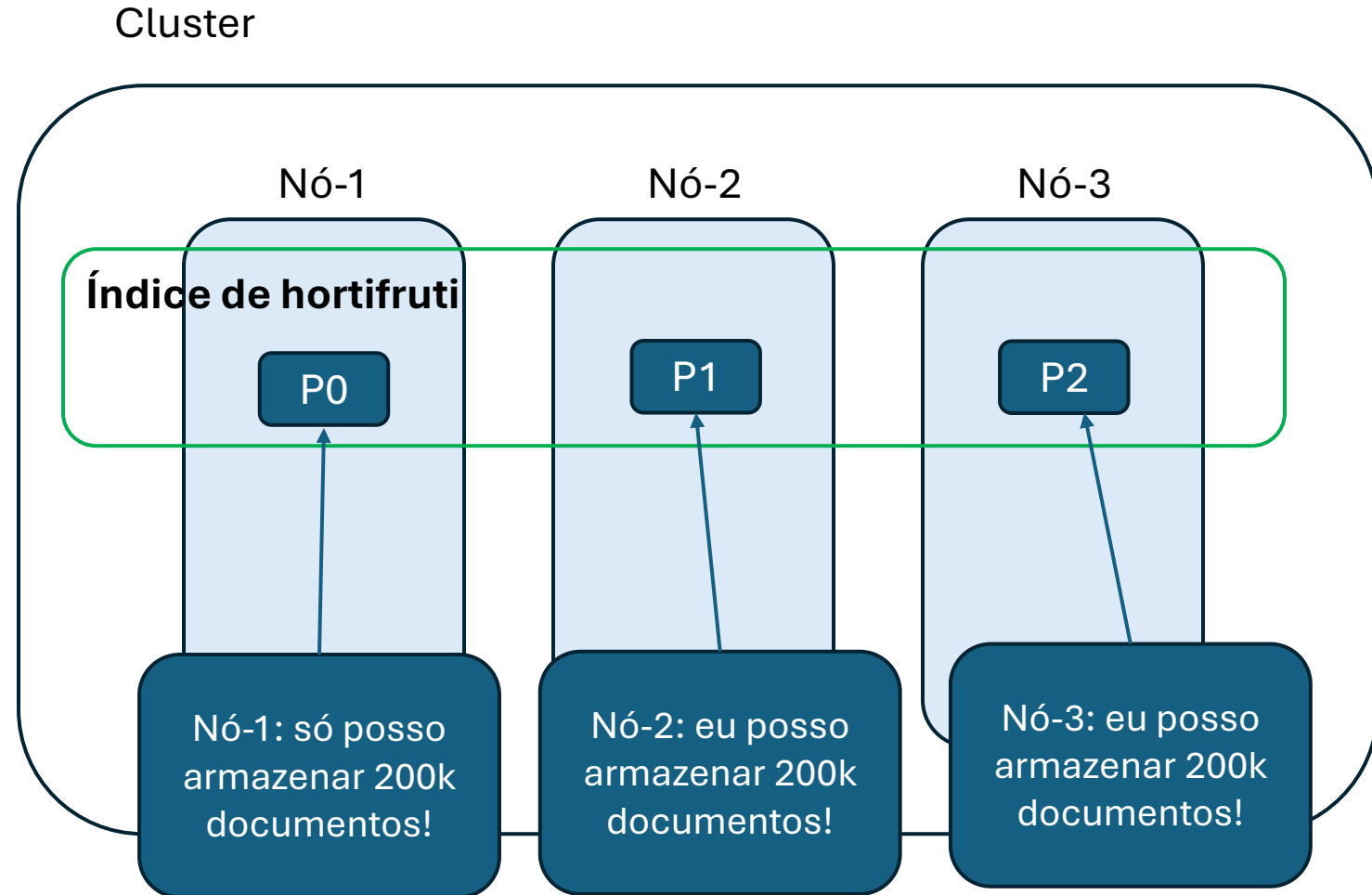


No momento da criação de um índice, um shard é criado automaticamente. É possível configurar o índice para distribuir os shards nos nós.

Shards

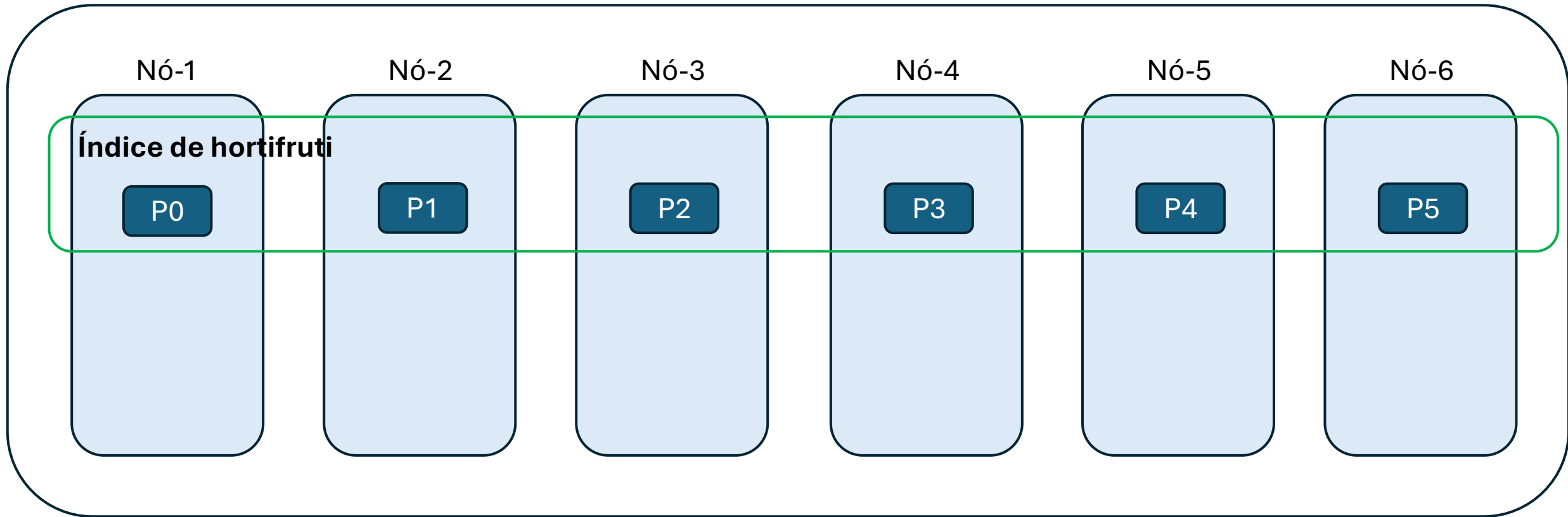
Cliente: quero armazenar 600k documentos no cluster

O número de documentos que um shard pode armazenar depende da capacidade de armazenamento do nó



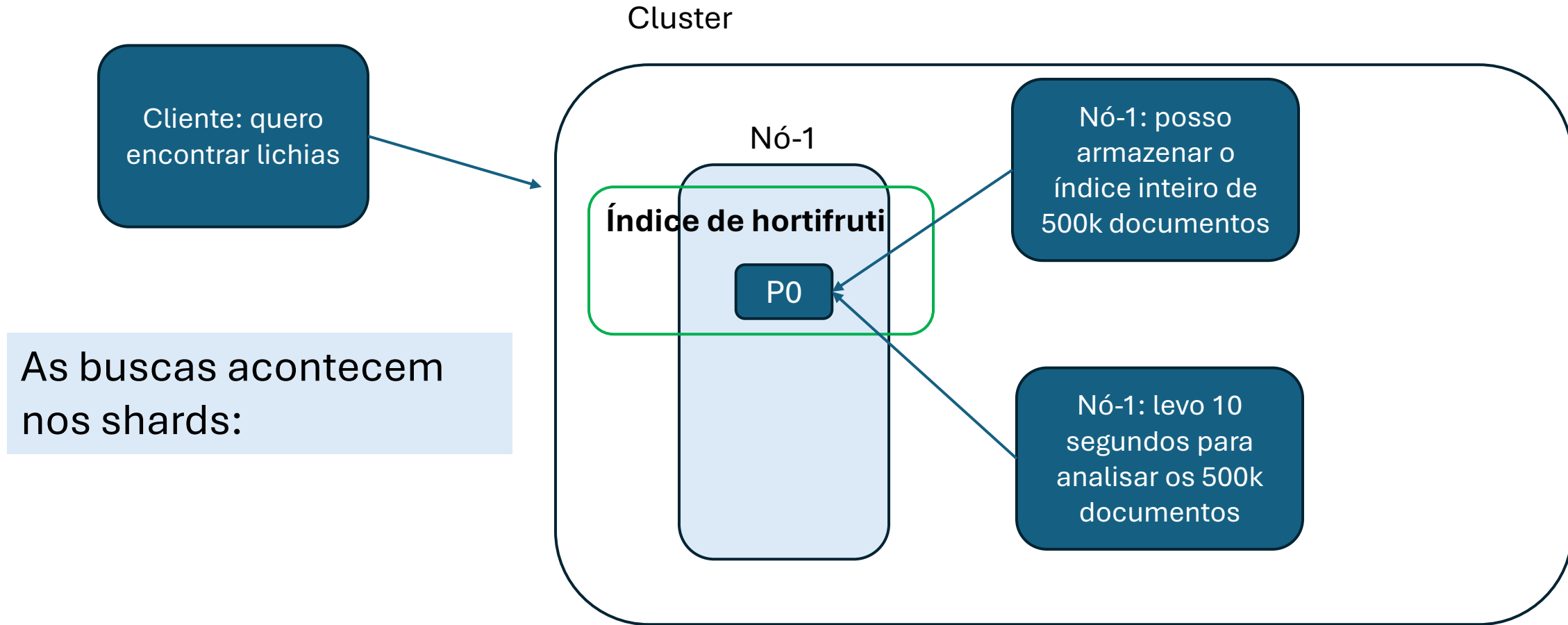
Shards

Cluster



   poss  vel adicionar mais shards conforme a necessidade.

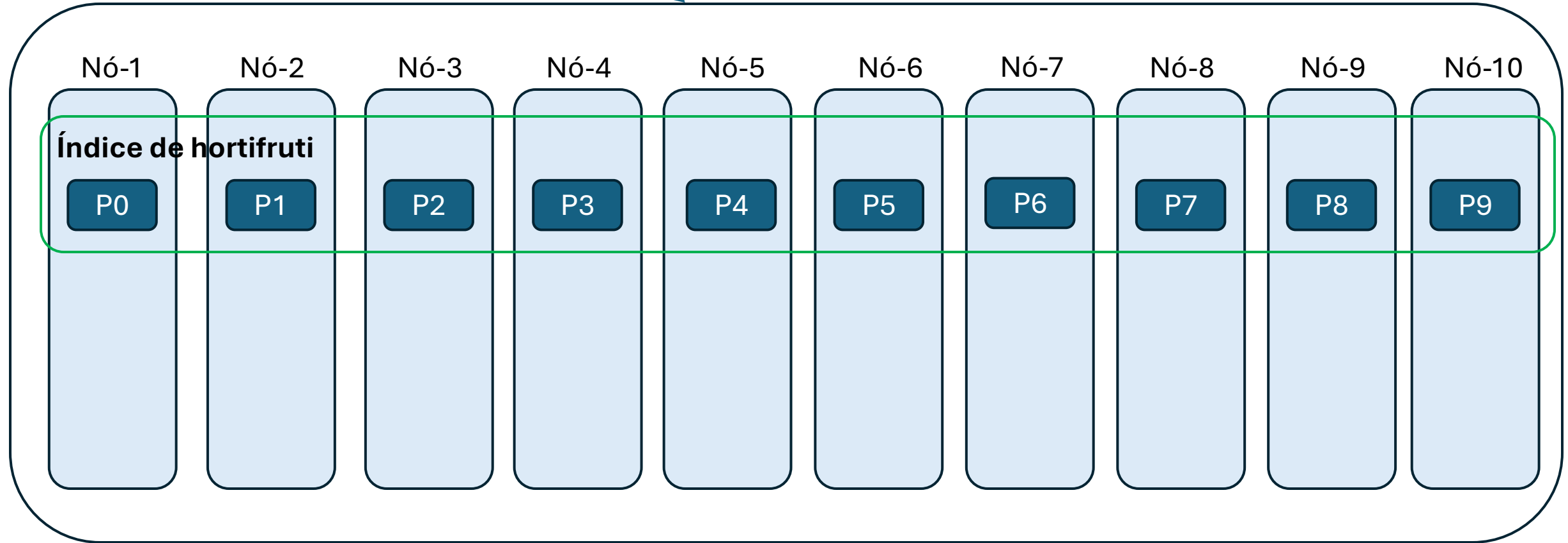
Shards



Shards

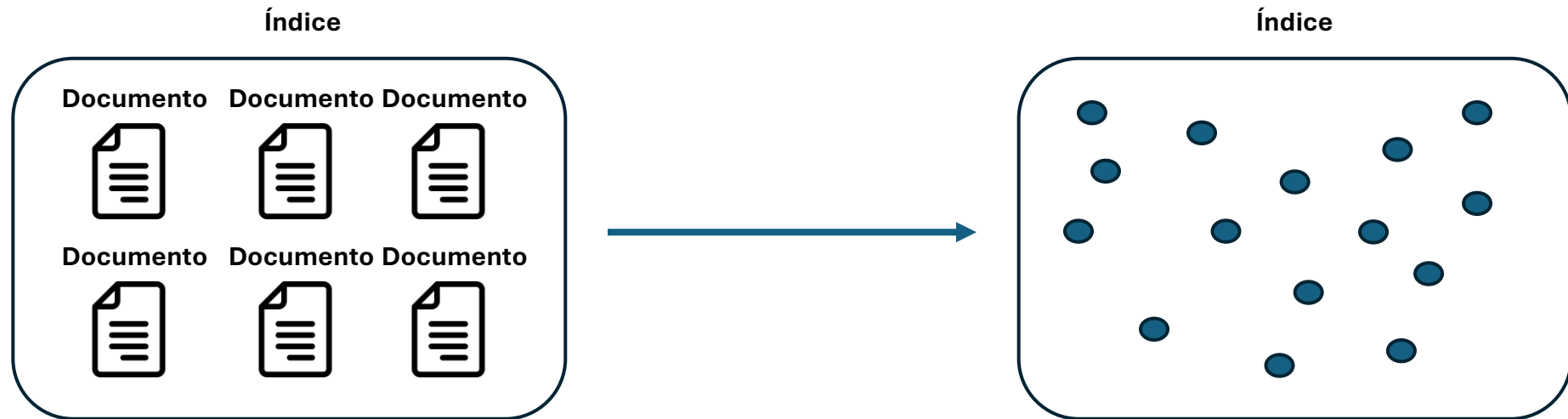
Cluster

Cluster: posso fazer a busca em paralelo, então levo apenas 1 segundo para buscar em 500k documentos



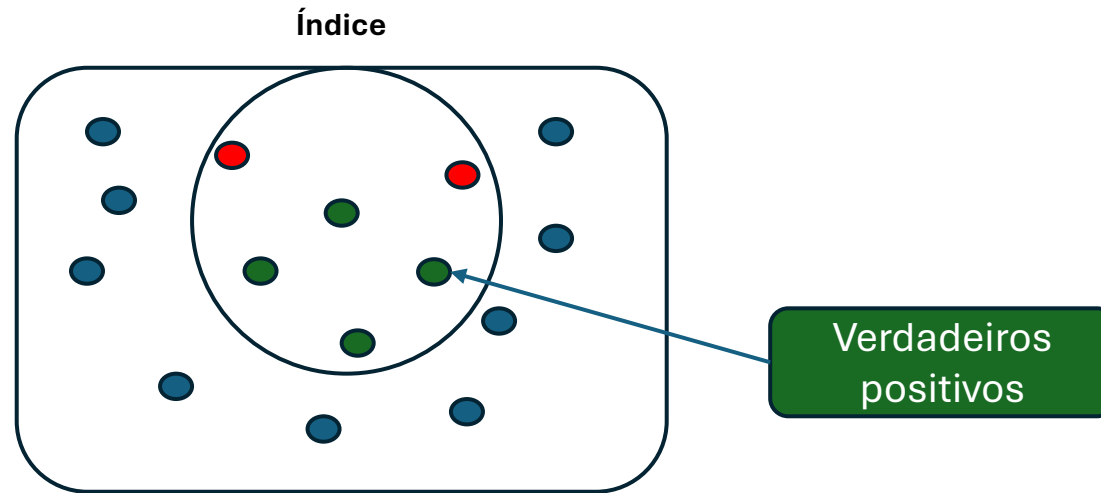
Se os documentos forem distribuídos por 10 nós, cada shard pode armazenar 50k documentos. Nesse caso, a busca levaria apenas 1 segundo.

Como acontecem as buscas?



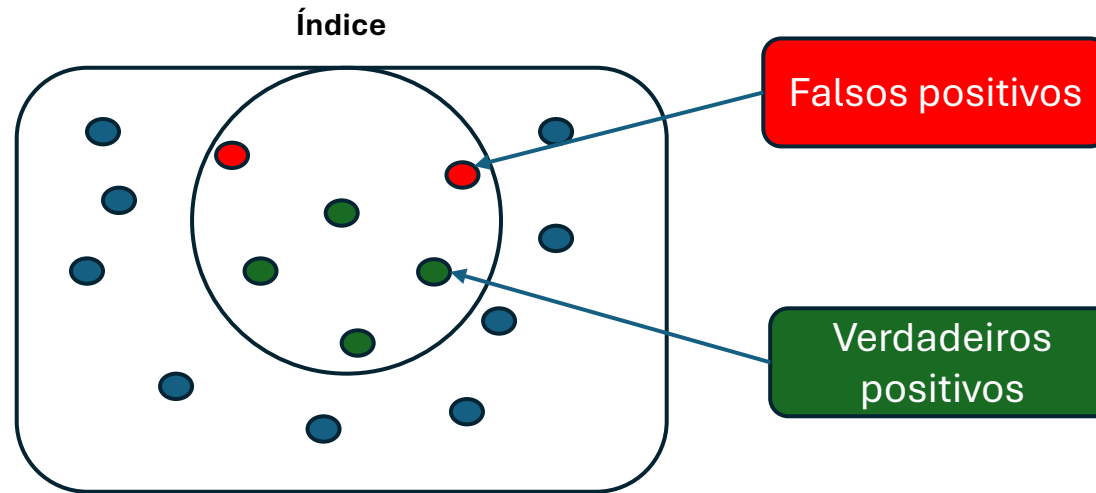
Vamos utilizar dois diagramas para mostrar a mesma coisa, mas para facilitar o entendimento das buscas do elasticsearch.

Verdadeiros positivos



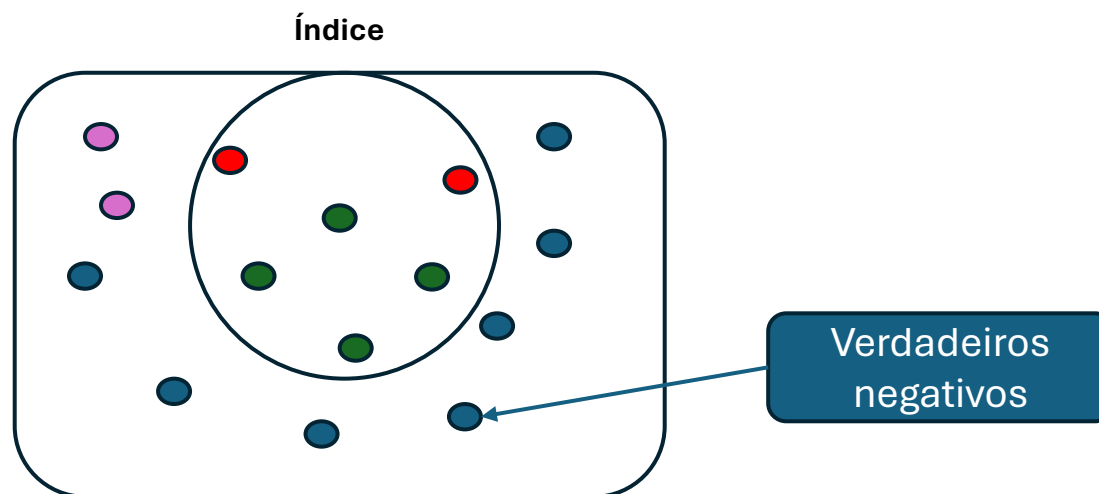
Verdadeiros positivos são documentos relevantes que são retornados para o usuário.

Falsos positivos



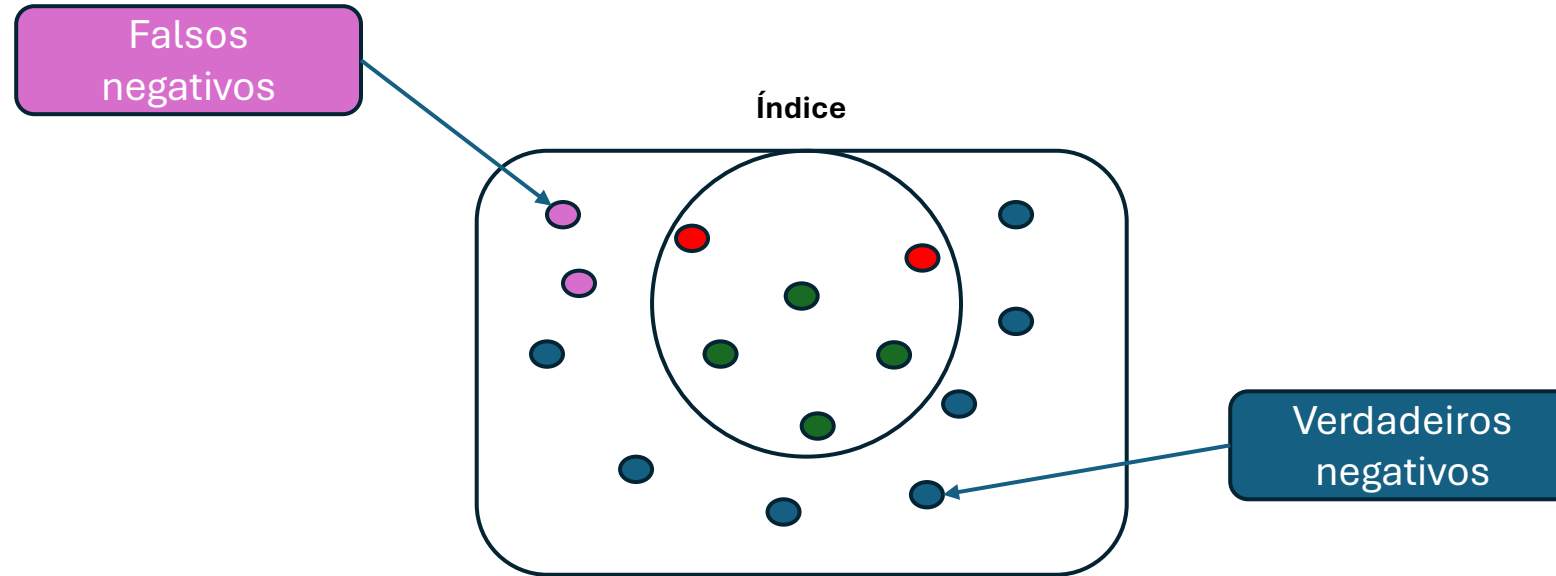
Falsos positivos são documentos irrelevantes que são retornados para o usuário.

Verdadeiros negativos



Verdadeiros negativos são documentos irrelevantes que não são retornados para o usuário.

Falsos negativos



Falsos negativos são documentos relevantes que não foram retornados para o usuário.

Precisão x Recall

$$\text{Precisão} = \frac{\text{Verdadeiros positivos}}{\text{Verdadeiros positivos} + \text{Falsos positivos}}$$

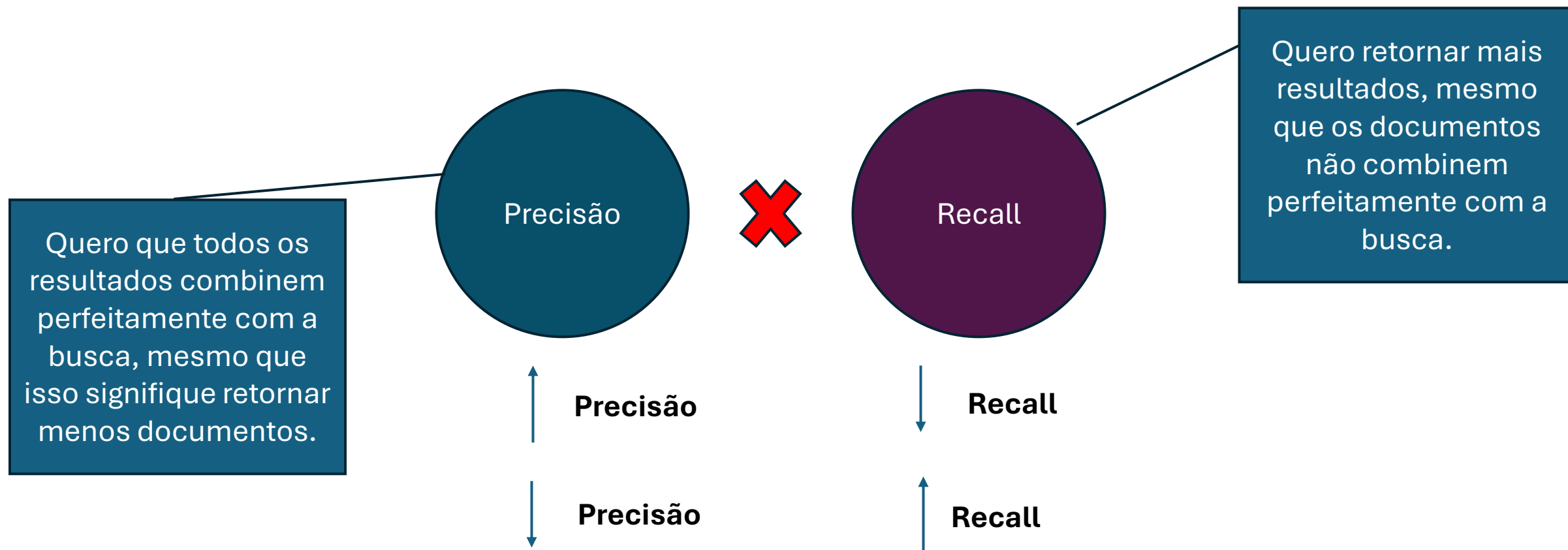
Que porção dos dados retornados é realmente relevante para a busca?

$$\text{Recall} = \frac{\text{Verdadeiros positivos}}{\text{Verdadeiros positivos} + \text{Falsos negativos}}$$

Que porção dos dados que são relevantes está sendo retornada nos resultados da busca?

Precisão x Recall

Precisão e Recall são inversamente proporcionais



Precisão e recall determinam quais documentos serão retornados, mas não a relevância entre eles.

Relevância do resultado

Existe uma pontuação para determinar qual documento é mais relevante e qual é menos relevante. Essa pontuação é calculada para cada documento retornado (hits).

Uma das formas de calcular essa pontuação é através do $TF \times IDF$

TF: Frequência do Termo

IDF: Frequência Inversa dos Documentos

🔍 | Como criar bons hábitos



Mais relevante (maior pontuação)

...

...

...

Menos relevante

...

...

...

Nada relevante (menor pontuação)

Frequência do Termo (TF)

Determina quantas vezes cada termo buscado aparece no documento.

The screenshot shows a search interface with a search bar containing the query "How to form good habits". The word "habits" is highlighted with a blue box. Below the search bar, two document results are displayed as JSON objects. The first result is for "Atomic Habits" by James Clear, and the second is for "The Mental Toughness Handbook" by Damon Zahariades. In both results, the word "habits" is underlined in the description. The first result has a TF of 4, and the second result has a TF of 1.

```
{
  "title": "Atomic Habits",
  "author": "James Clear",
  "category": "self-help",
  "description": "No matter your goals, Atomic Habits offers a proven framework for improving every day. James clear, ... habits...habits ...habits" TF= 4
}

{
  "title": "The Mental Toughness Handbook",
  "author": "Damon Zahariades",
  "category": "self-help",
  "description": "Imagine boldly facing any challenge that comes your way... 5 daily habits you must embrace to strengthen your mind and harden your resolve. Why willpower and motivation are unreliable..." TF= 1
}
```

Se os termos buscados são encontrados com uma alta frequência em um documento, ele é considerado mais relevante para a busca.

Frequência Inversa dos Documentos (IDF)

🔍 | Como criar bons hábitos

Como criar uma conversa em grupo



Bons pratos com frango



Como criar uma banda



Bons lugares para visitar



Manual de bons hábitos



Bons hábitos envolvem dedicação



Hits

Podem ter alguns dos termos, mas não tem nada a ver com criação de bons hábitos.

O IDF diminui o peso de termos que ocorrem muito frequentemente no conjunto de documentos e aumenta o peso de termos que ocorrem raramente.

Prática com o Elasticsearch

Vamos nos conectar a uma máquina do laboratório e, usando o Kibana, fazer consultas ao Elasticsearch.

Para isso, no navegador do computador do laboratório, digite:

10.8.1.38:5601

A página do Kibana deverá aparecer.

Usaremos os comandos da página do Github da aula para realizar as buscas: <https://bit.ly/aula-elastic>

Introdução ao Elasticsearch

