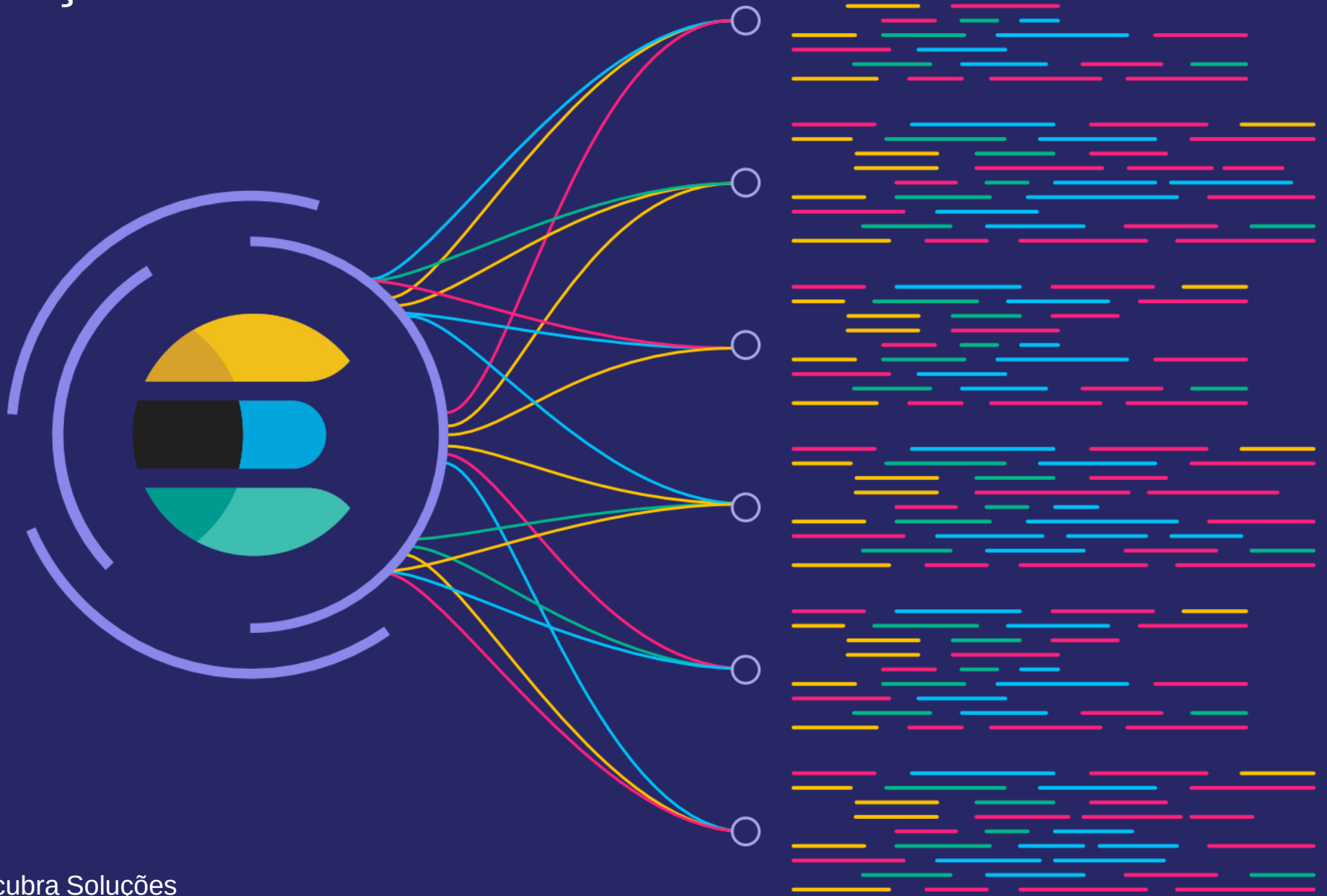
Introdução ao Elasticsearch



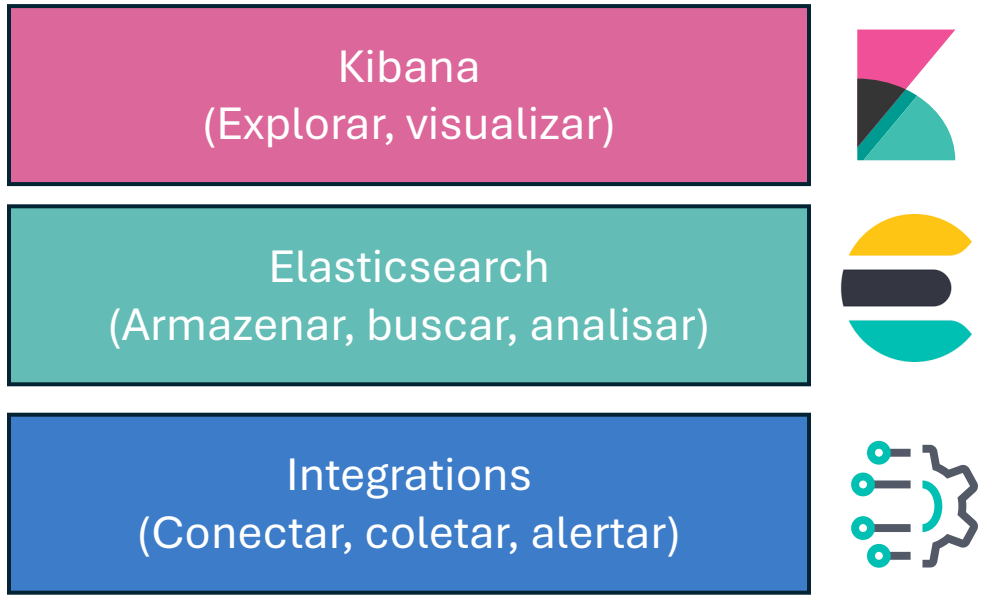
Flávio Mota
COO - Descubra Soluções

O que é o Elasticsearch?

elastic stack

Elastic Stack

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



O que é o Elasticsearch?

Elasticsearch
(Armazenar, buscar, analisar)

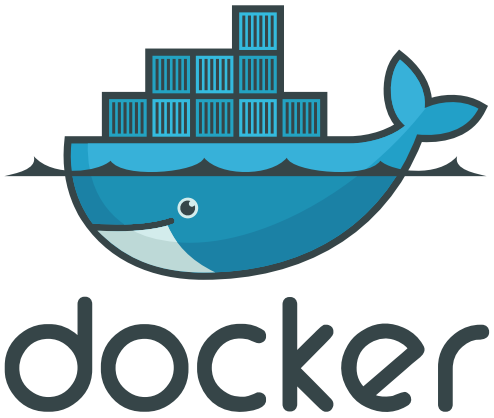
Elasticsearch é um mecanismo de busca em tempo real, baseado em documentos, que permite o armazenamento, a pesquisa e a análise de grandes volumes de dados rapidamente.

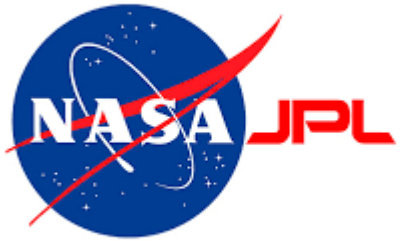
Mecanismos de buscas

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.

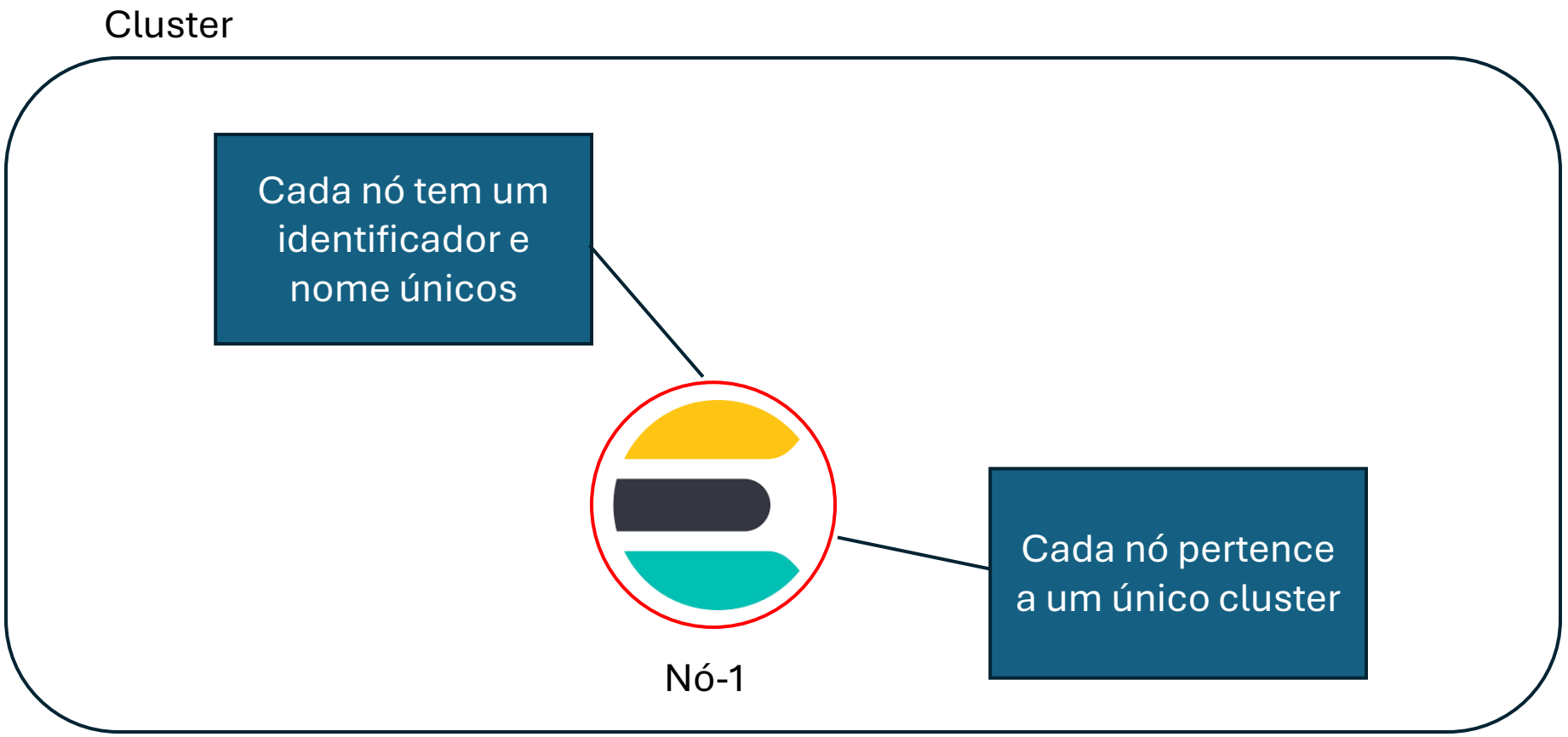
Boa experiência de busca = Obter resultados relevantes de forma rápida, independente da escala

Aplicações que utilizam o Elasticsearch

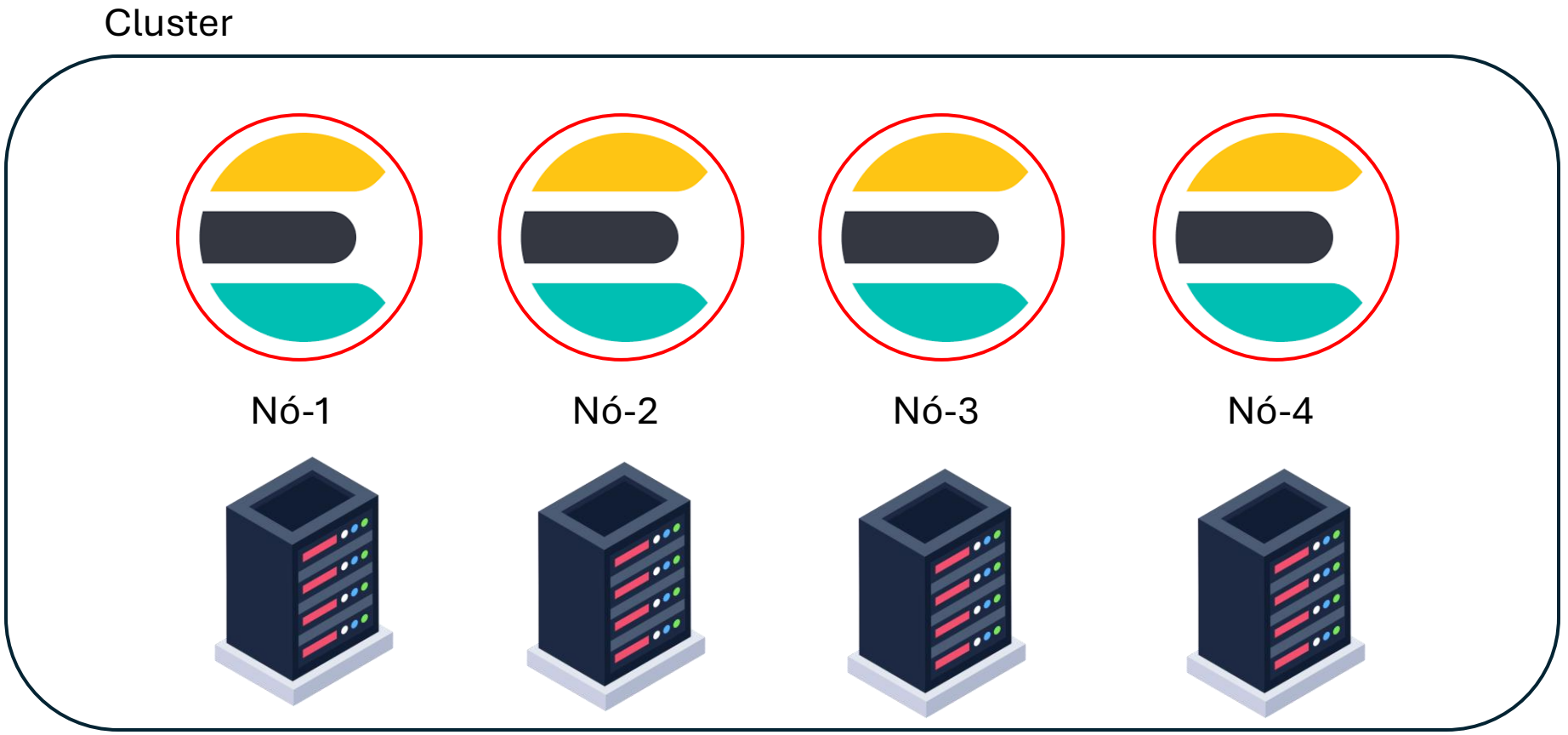




Arquitetura do Elasticsearch



Arquitetura do Elasticsearch



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

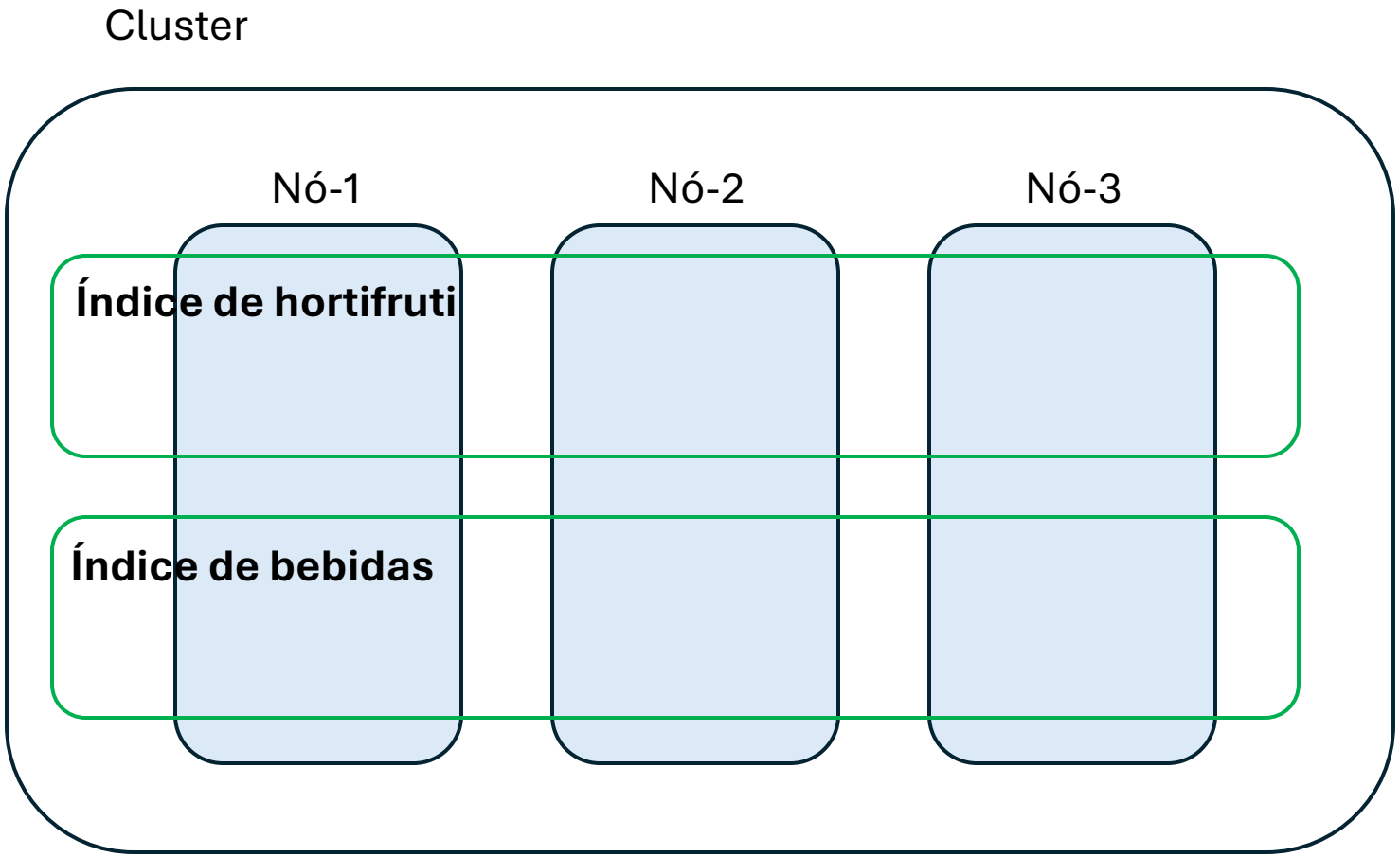
```
{
    "nome":"Banana Prata",
    "categoria": "Frutas",
    "preco": "R$ 6,37"
}
```

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

Índice de bebidas

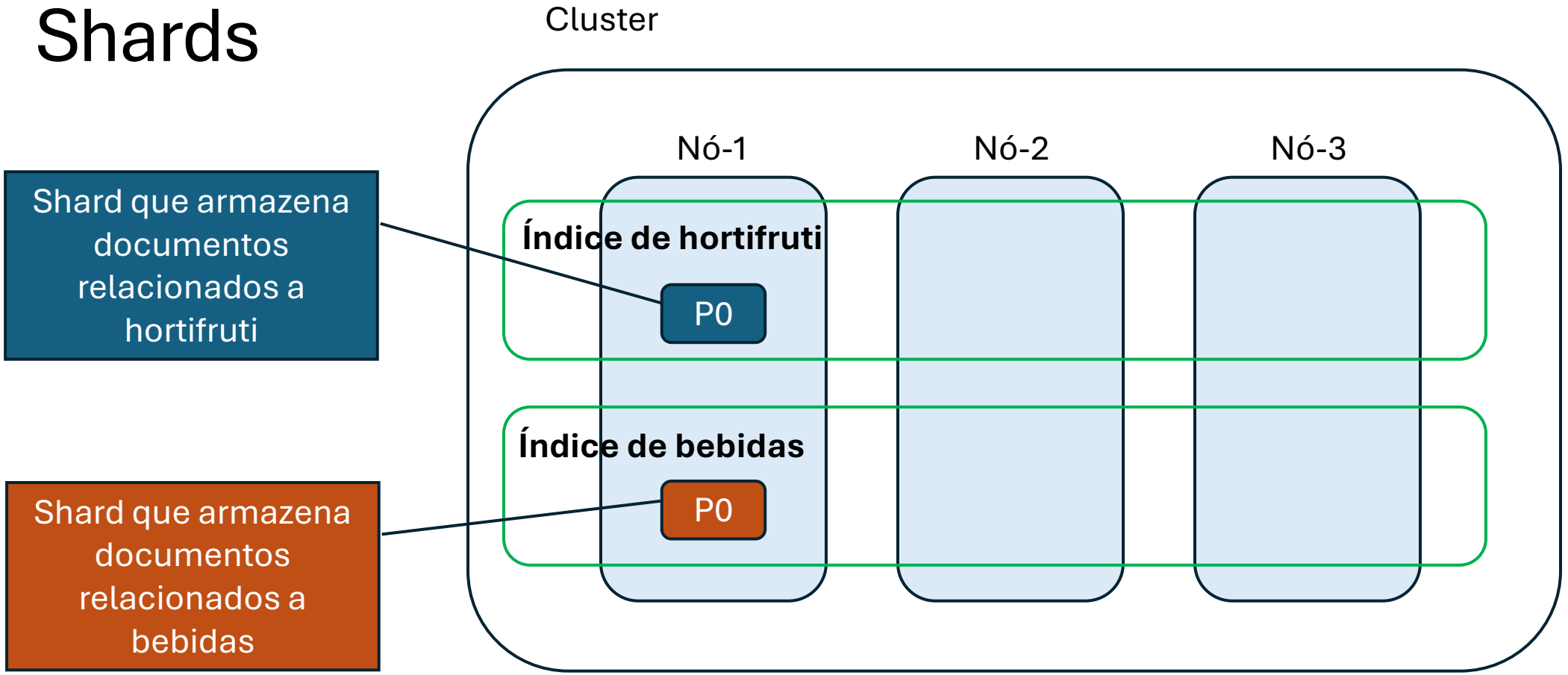
```
{
    "nome": "Spaten",
    "categoria": "Cerveja",
    "preco": "R$ 4,50",
    "tipo": "Lata"
}
```

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



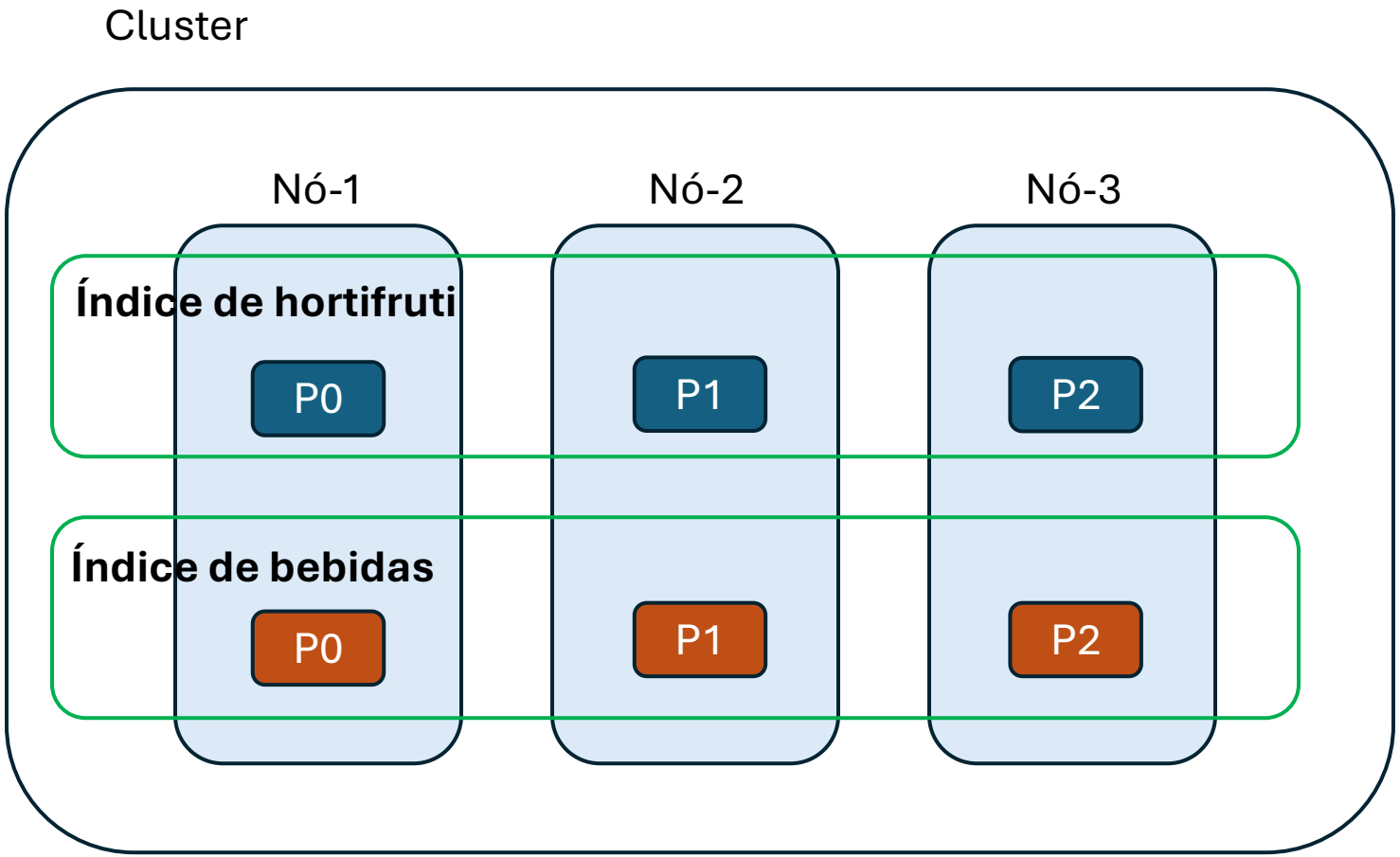
O índice não armazena os documentos. Ele é um recurso virtual que mantém um controle de onde os documentos estão armazenados!

Shards



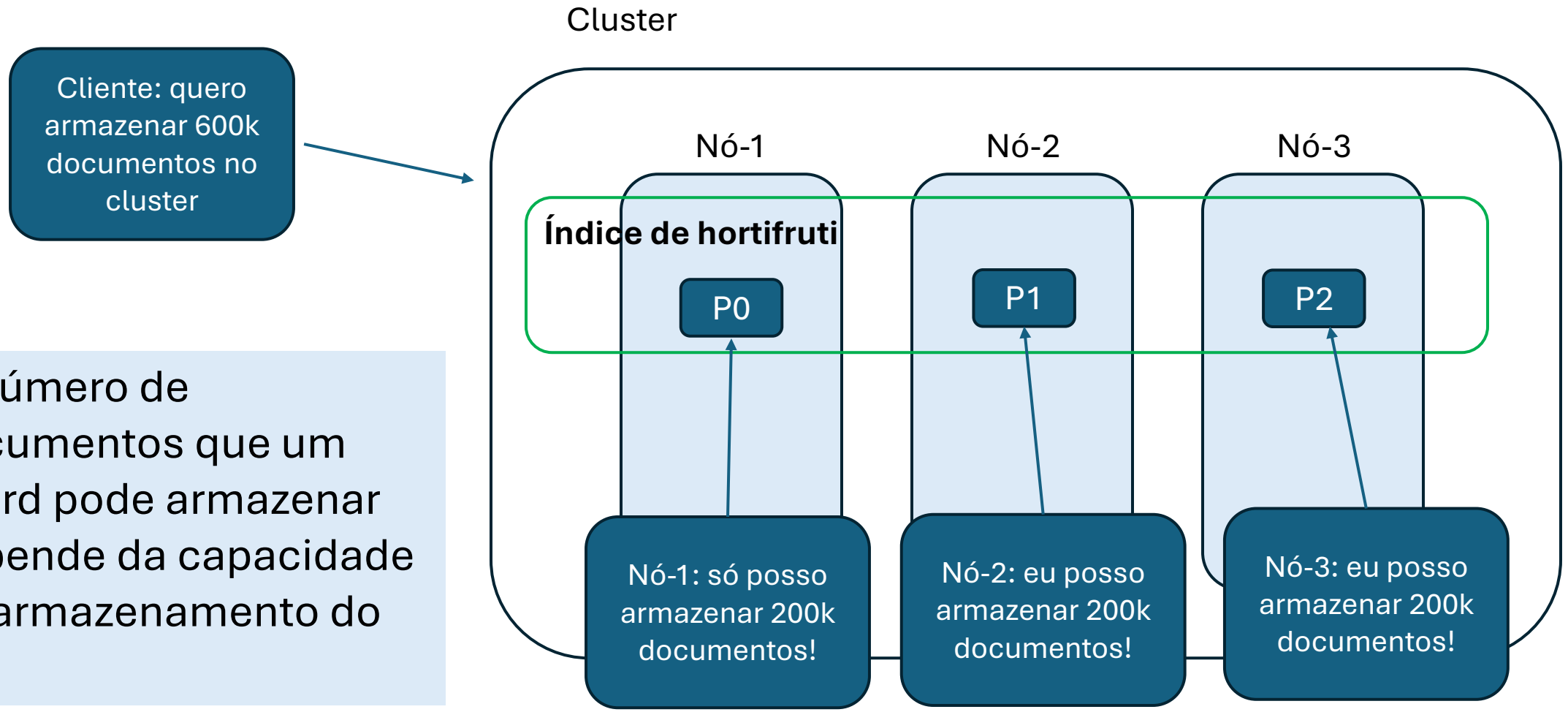
Shard é onde os dados são armazenados e a busca é realizada.

Shards



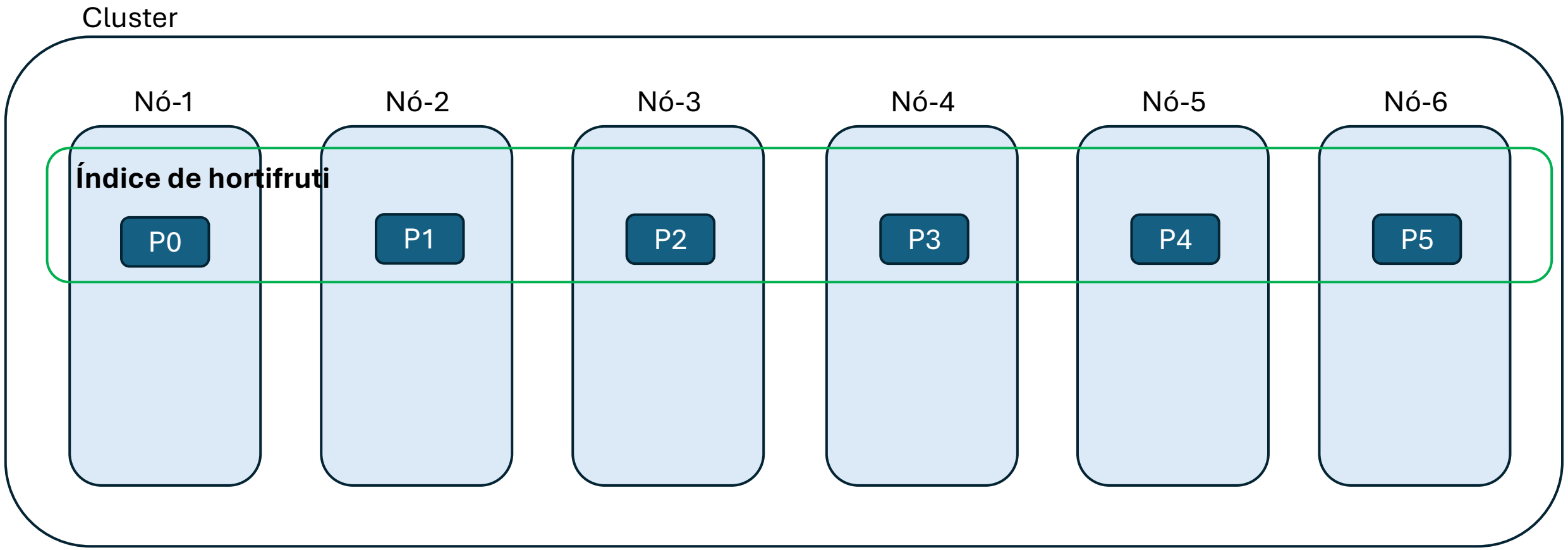
No momento da criação de um índice, um shard é criado automaticamente. É possível configurar o índice para distribuir os shards nos nós.

Shards



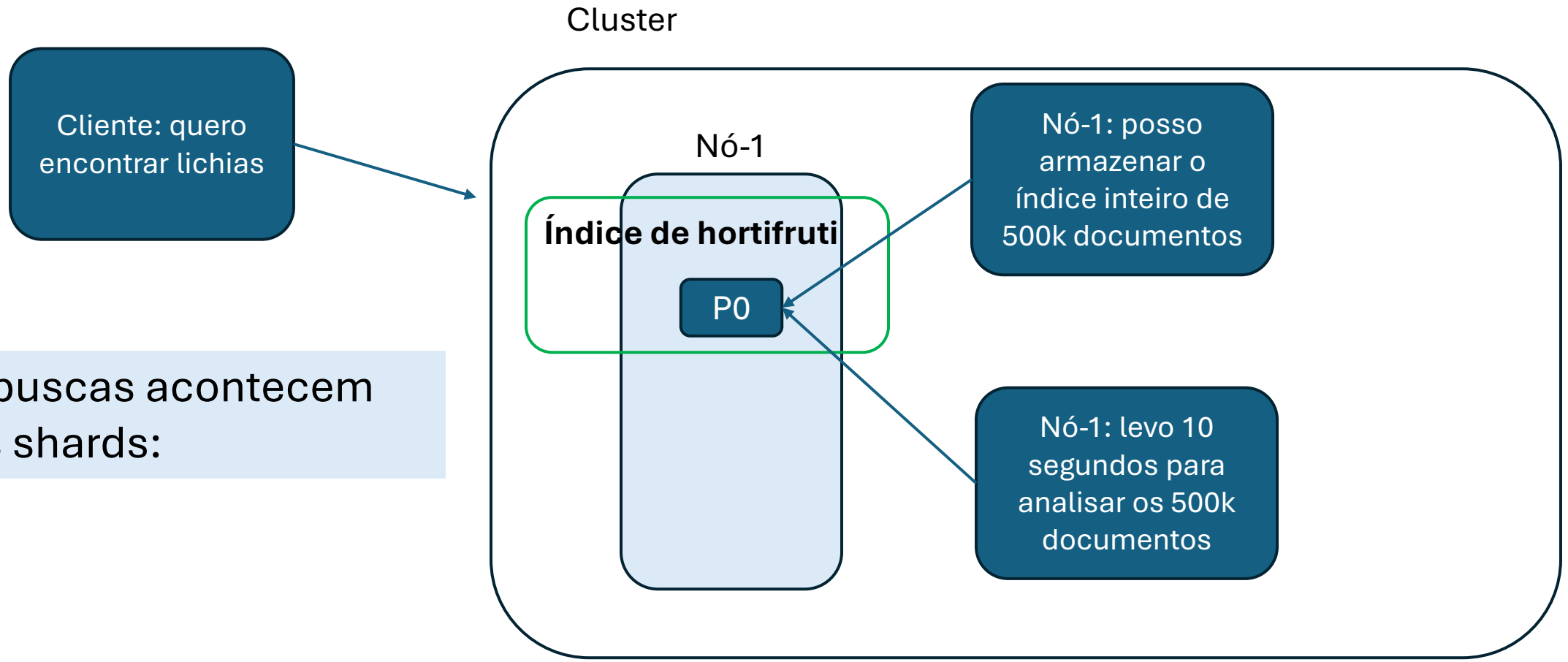
O número de documentos que um shard pode armazenar depende da capacidade de armazenamento do nó

Shards



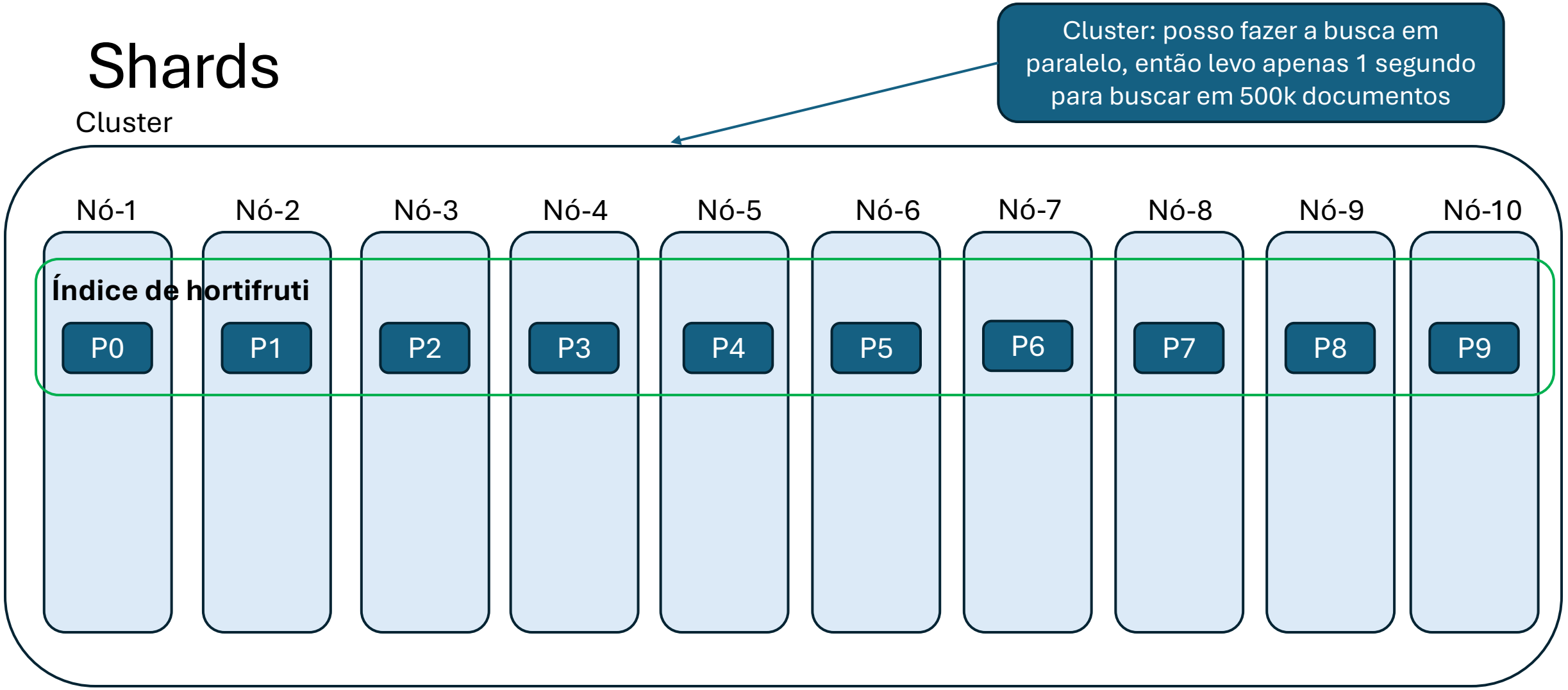
É possível adicionar mais shards conforme a necessidade.

Shards



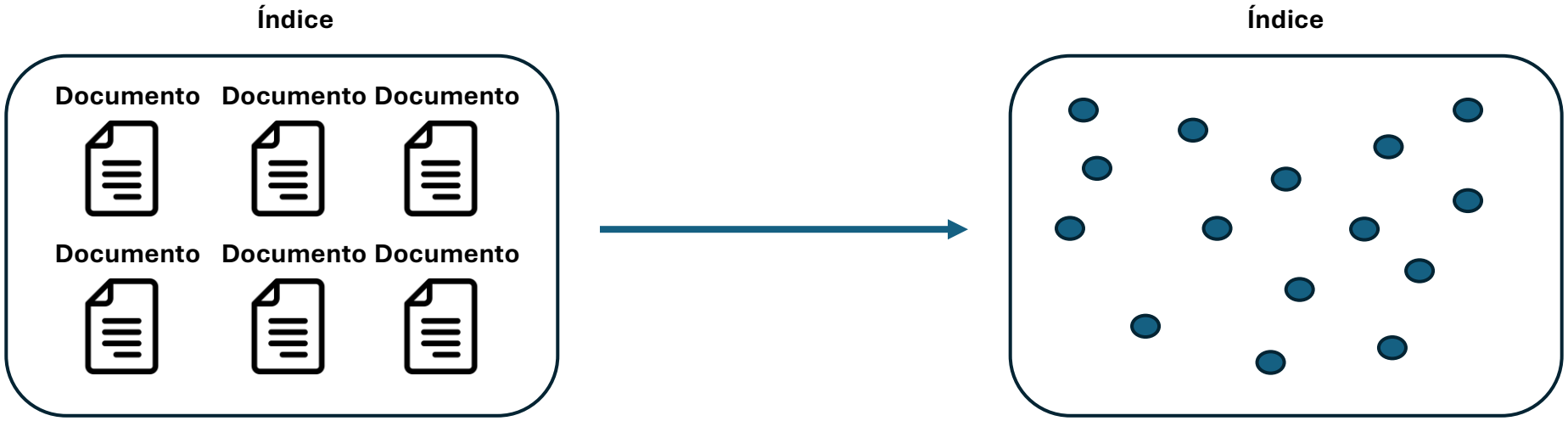
As buscas acontecem nos shards:

Shards



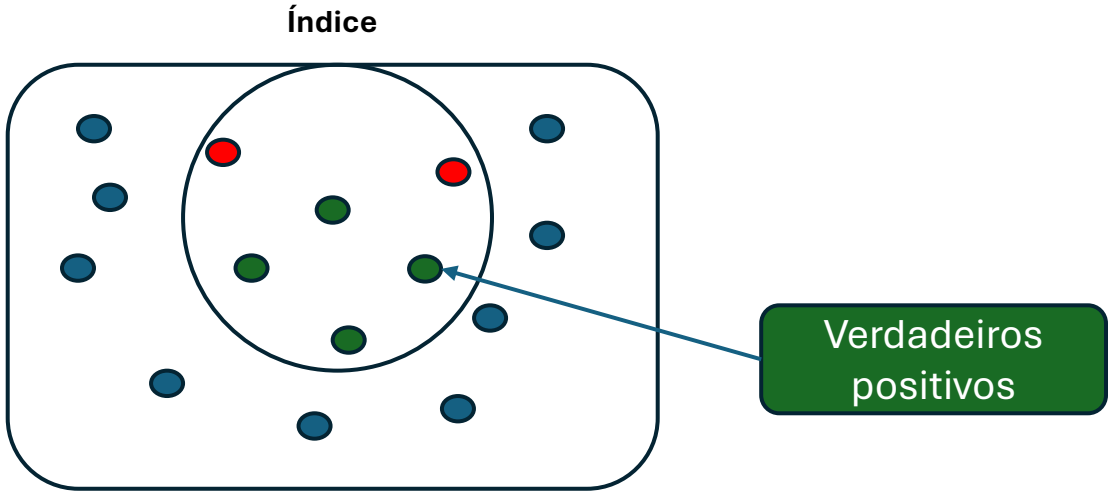
Se os documentos forem distribuídos por 10 nós, cada shard pode armazenar 50k documentos. Nesse caso, a busca levaria apenas 1 segundo.

Como acontecem as buscas?



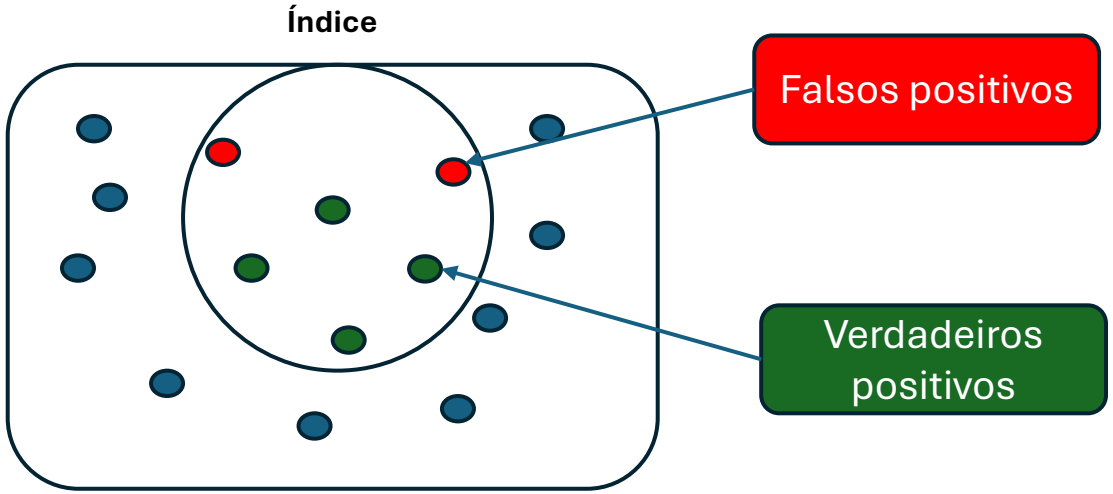
Vamos utilizar dois diagramas para mostrar a mesma coisa, mas para facilitar o entendimento das buscas do elasticsearch.

Verdadeiros positivos



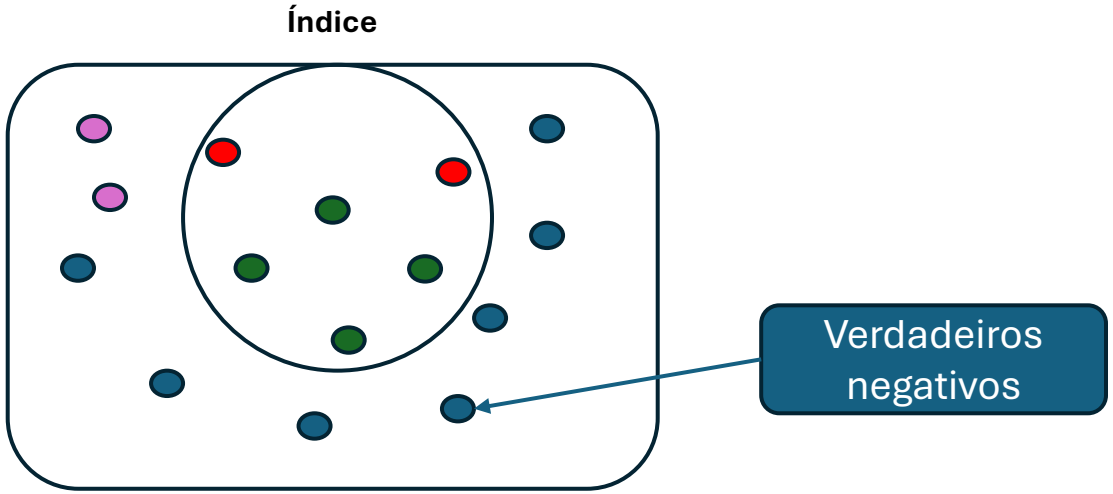
Verdadeiros positivos são documentos relevantes que são retornados para o usuário.

Falsos positivos



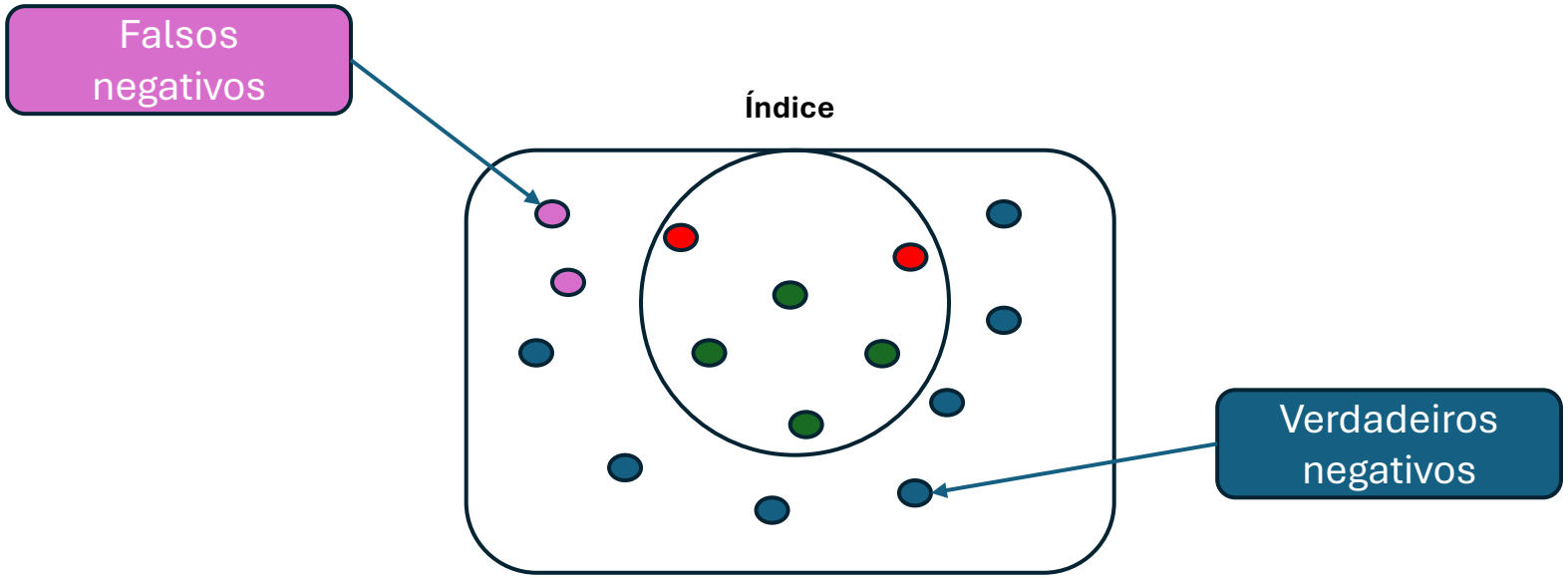
Falsos positivos são documentos irrelevantes que são retornados para o usuário.

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

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

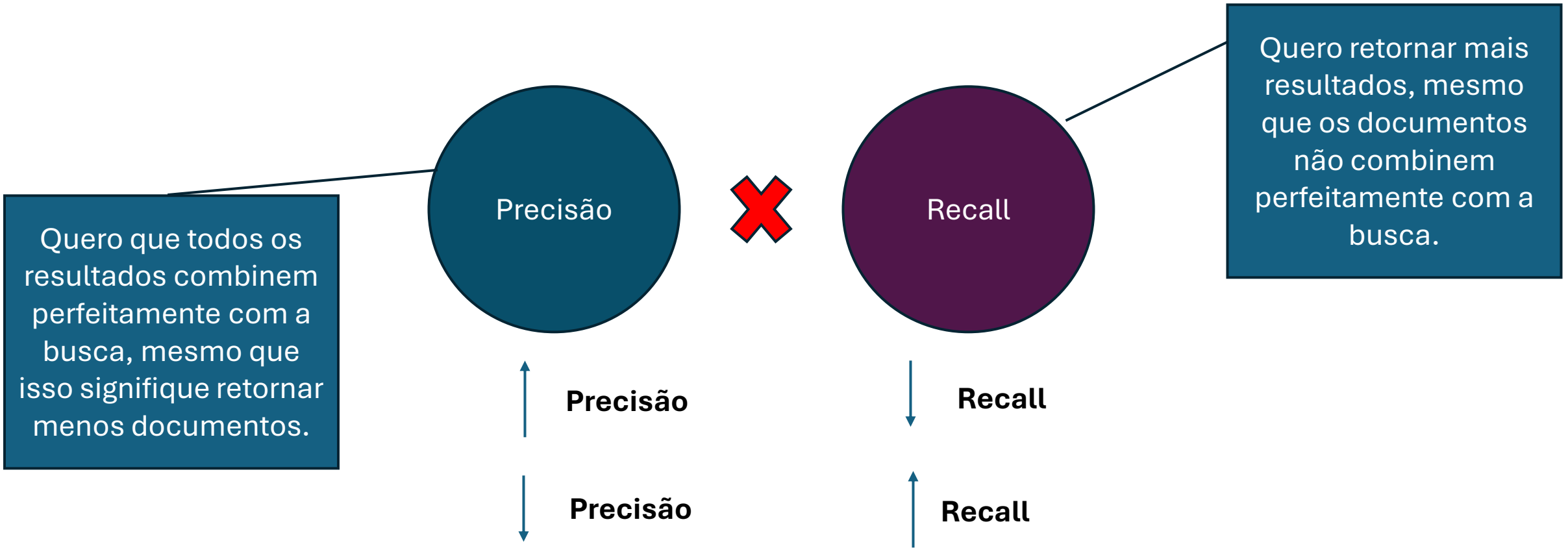
Que porção dos dados retornados é realmente relevante para a busca?

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

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

Precisão x Recall

Precisão e Recall são inversamente proporcionais



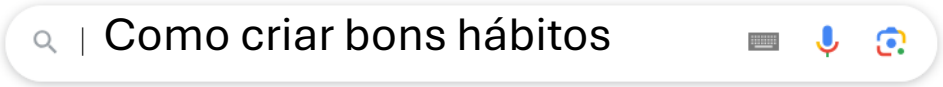
Precisão e recall determinam quais documentos serão retornados, mas não a relevância entre eles.

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 TFxIDF

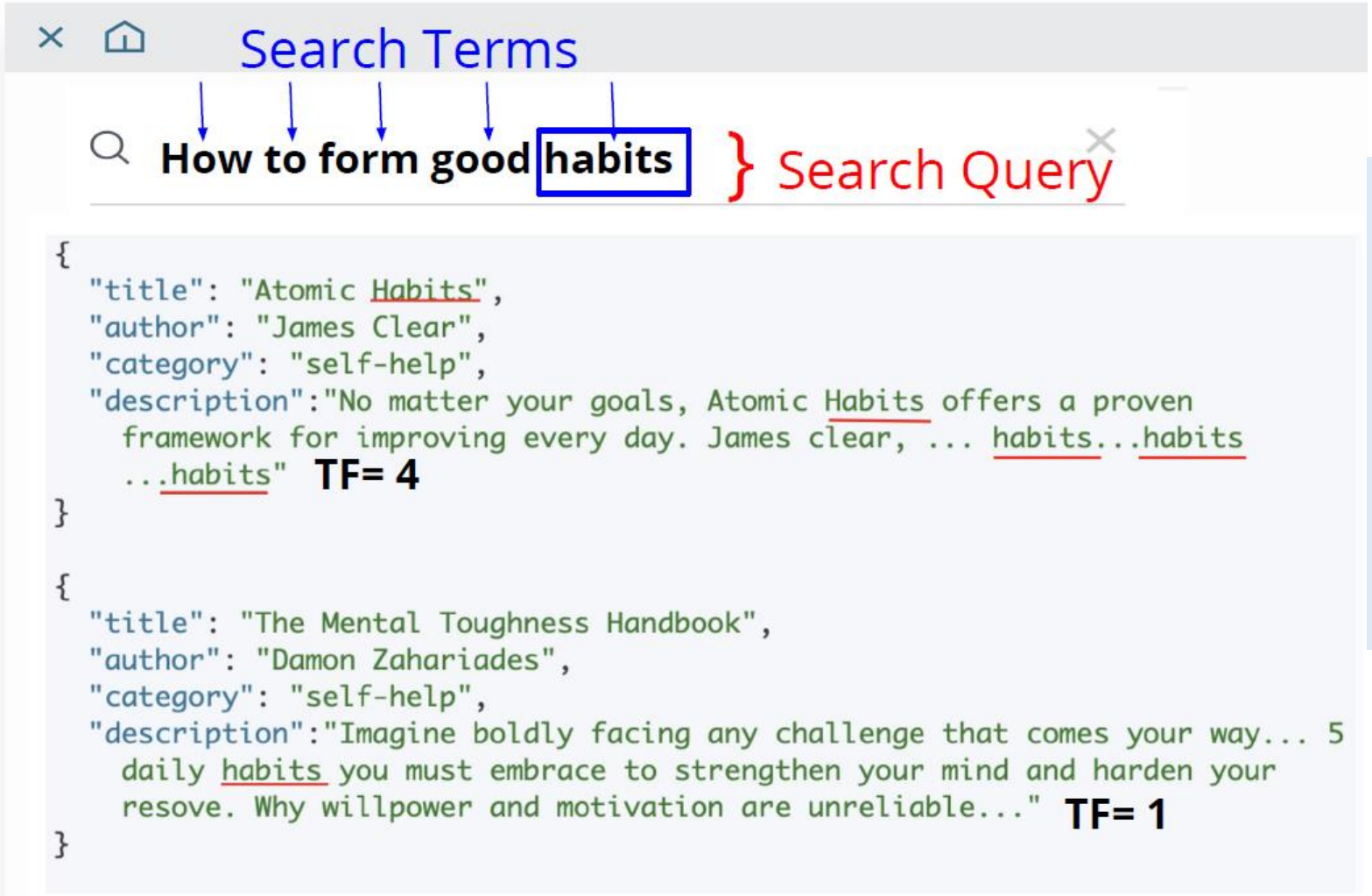
TF: Frequência do Termo
IDF: Frequência Inversa dos Documentos



Mais relevante (maior pontuação)
...
...
...
Menos relevante
...
...
...
Nada relevante (menor pontuação

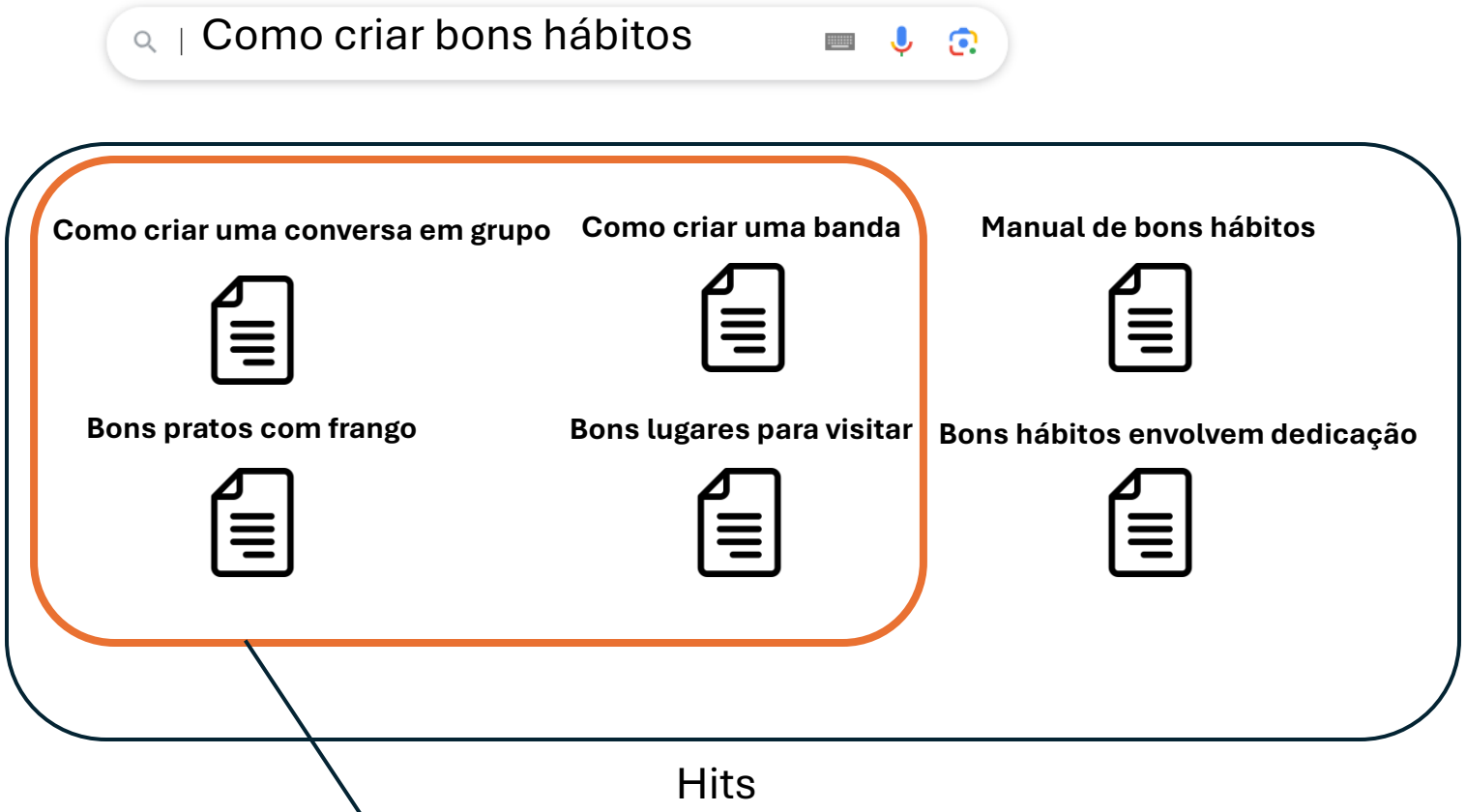
Frequência do Termo (TF)

Determina quantas vezes cada termo buscado aparece no documento.



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)



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.

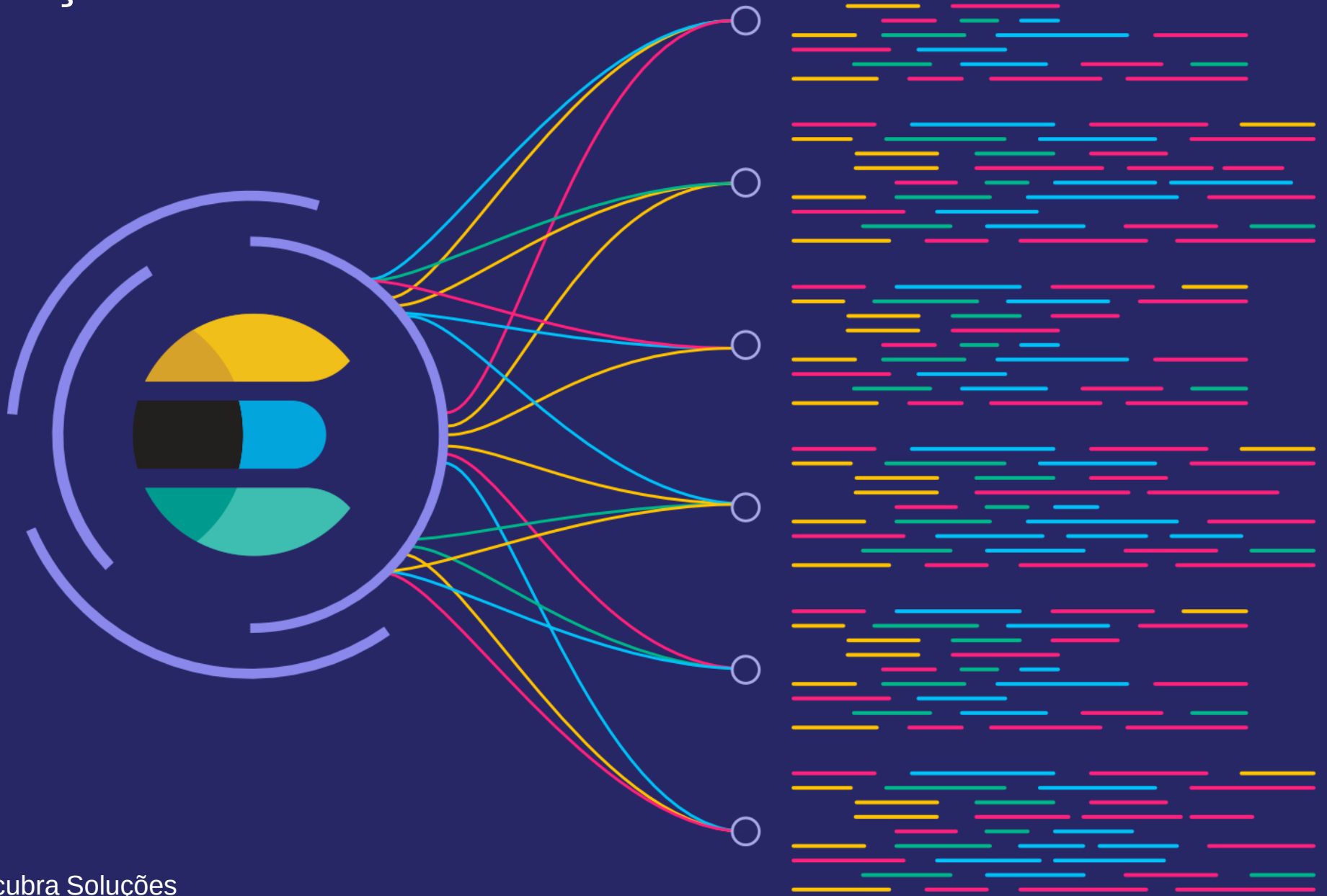
Para isso, no navegador do computador do laboratório, digite:

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



Flávio Mota
COO - Descubra Soluções