

Atributos de Qualidade em Elasticsearch

Desempenho e Segurança para Busca Vetorial em Sistemas Distribuídos





Desafio Principal

Como garantir performance em buscas vetoriais de larga escala?

Escala de Dados

Milhões ou bilhões de vetores em tempo real

Requisitos

Baixa latência com alta precisão

A busca vetorial kNN exata compara cada query com todos os vetores no dataset, criando um gargalo computacional crítico em sistemas distribuídos.

O Problema da Busca kNN Exata

Busca Vetorial (kNN)


Encontra vetores mais similares a uma consulta em grandes conjuntos de dados multidimensionais.

- Comparação por distância euclidiana
- Análise de similaridade vetorial
- Processamento sequencial completo

Limitações de Escala

A busca exata torna-se computacionalmente inviável com crescimento dos dados.

- Alta latência em milhões de vetores
- Consumo intensivo de CPU
- Impacto no throughput do sistema



Approximate Nearest Neighbor (ANN)

A solução inteligente para performance em larga escala

1 Trade-off Estratégico

Sacrifica pequeno percentual de precisão (1-5%) para ganhos exponenciais de velocidade

2 Algoritmo HNSW

Elasticsearch implementa Hierarchical Navigable Small World para buscas ANN nativas

3 Performance Superior

Ordens de magnitude mais rápido que busca exata em datasets grandes

Arquitetura do HNSW

Estrutura hierárquica multi-camadas para navegação eficiente

Estrutura Hierárquica

Vetores organizados em camadas - topo esperso, base densa com todos os nós

Navegação Rápida

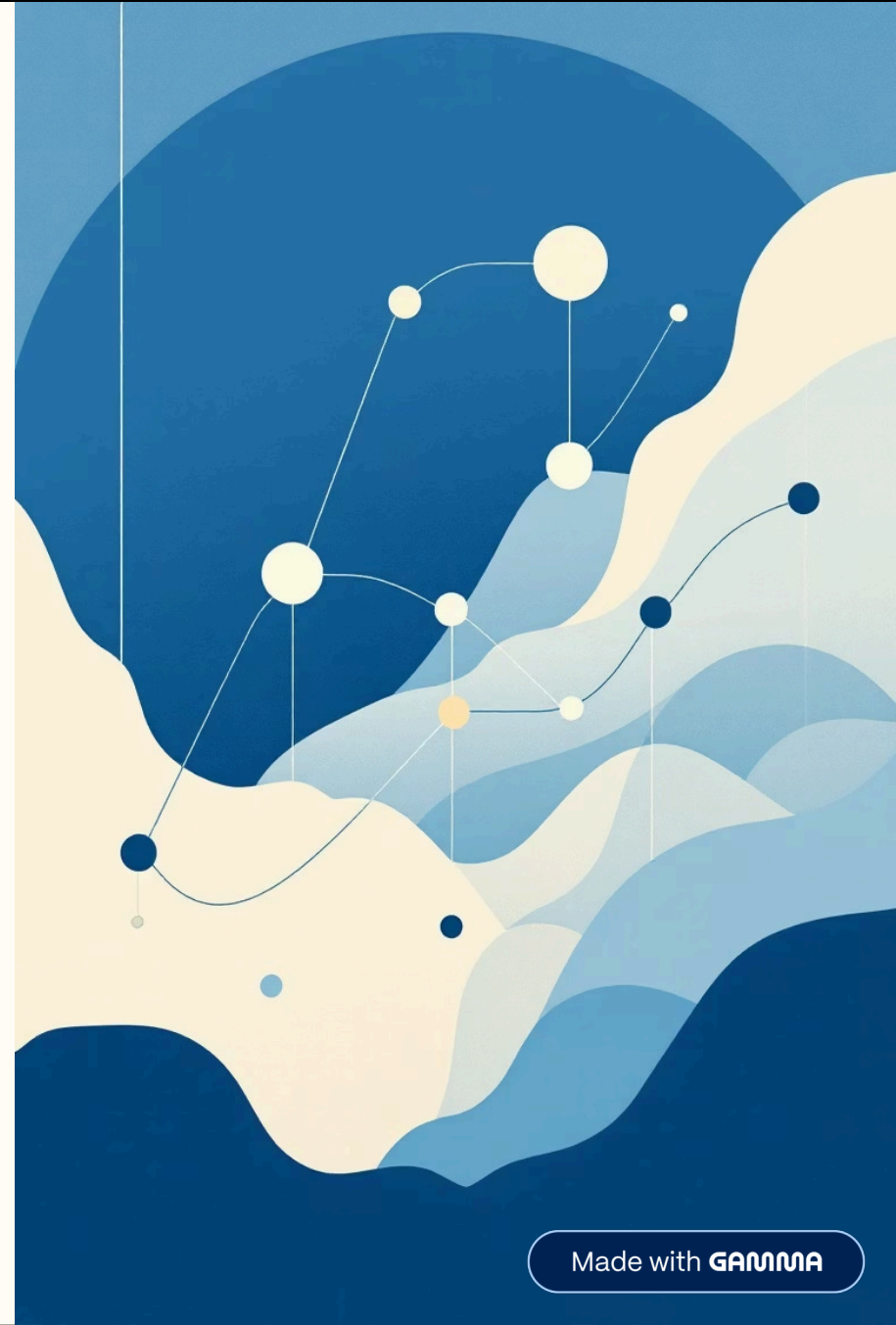
Algoritmo navega camada superior encontrando vizinho mais próximo do nível

Descida Refinada

Usa nó encontrado como ponto de partida para busca na camada inferior

Resultado Otimizado

Processo iterativo até camada base, descartando grandes partes do grafo



Comparativo: kNN Exato vs ANN

Característica	kNN Exato (script_score)	ANN com HNSW (knn search)
Precisão	100% (garantido)	95-99% (configurável)
Performance	Lento, alto custo de CPU	Extremamente rápido
Uso de Memória	Baixo	Alto (grafo HNSW em memória)
Caso de Uso	Datasets pequenos, precisão crítica	Tempo real, larga escala

Segurança em Camadas

Proteção abrangente para sistemas distribuídos

Autenticação

Quem é você?

Verificação de identidade de usuários e serviços



Validação

Como testar segurança?

Pentests e auditoria contínua



Autorização

O que você pode fazer?

Controle granular de permissões e acessos



Criptografia

Como proteger dados?

Segurança em trânsito e em repouso





Autenticação: Primeira Linha de Defesa

Processo de verificação de identidade antes do acesso ao cluster

Realm Nativo

Usuários e senhas gerenciados diretamente pelo Elasticsearch

- Controle total interno
- Configuração simplificada
- Ideal para ambientes isolados

LDAP/Active Directory

Integração com serviços de diretório corporativo

- Centralização de identidades
- Sincronização automática
- Políticas de senha existentes

SAML/OpenID Connect

Single Sign-On para experiência unificada

- Autenticação federada
- Redução de credenciais
- Integração com SSO corporativo

RBAC: Autorização Granular

Role-Based Access Control para controle preciso de permissões



Privilégios

Ações granulares específicas como read, write, cluster:monitor, indices:admin



Roles (Papéis)

Coleções de privilégios aplicados a índices específicos ou cluster inteiro

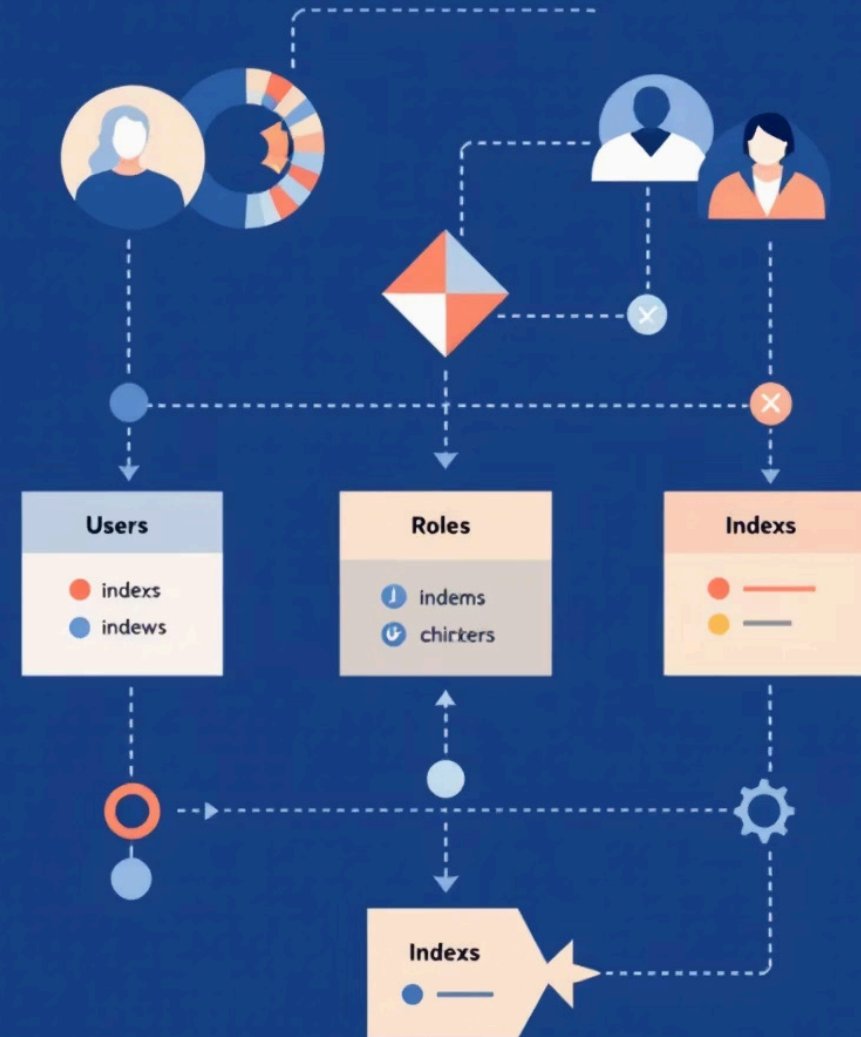


Mapeamento

Associação flexível de usuários a múltiplos papéis conforme necessidade



Princípio do Menor Privilégio: Usuários recebem apenas permissões mínimas necessárias para suas funções



Criptografia: Proteção de Dados

Em Trânsito (TLS/SSL)

Protege comunicação entre nós do cluster e clientes externos

- Certificados X.509
- Comunicação inter-node
- API REST segura
- Transport layer encryption



Em Repouso

Criptografa dados fisicamente armazenados em disco

- Encryption at rest
- Proteção de shards
- Logs e snapshots
- Chaves de criptografia seguras



Resultado: Confidencialidade garantida em todas as camadas de armazenamento e transmissão