Search Guard Secures the Retrieval Layer of AI Applications on Elasticsearch

Anja Glauch · 2026-11-05



Search Guard Secures the Retrieval Layer of AI Applications on Elasticsearch

Document- and field-level security decide what an AI retrieval pipeline may return, per user and attribute. Access is logged. No extra AI security tool needed.

Berlin, Germany, 5 November 2026. Elasticsearch has become a standard component in retrieval-augmented generation (RAG): it holds the documents and vector indices that AI assistants search before they answer. That makes it the place where an AI application touches sensitive data, and the place where access decisions have to be made. Search Guard, a product of floragunn GmbH in Berlin, applies its existing access control, encryption and audit logging to these workloads, so that an assistant can only retrieve what the person asking is allowed to see.

The Risk Sits in Retrieval

A RAG system answers from whatever its search layer returns. If the retrieval step runs with broad service permissions, an assistant can surface HR records, customer data or internal reports to users who would never be given access to those documents directly. The language model is rarely where this goes wrong. The query to Elasticsearch is.

Search Guard enforces document-level and field-level security on every request that reaches Elasticsearch. For an AI service, the key is that retrieval runs under the identity of the person asking rather than a broad service account: either by passing the user's token through to Elasticsearch, or through Search Guard's user impersonation, which lets a trusted service act on behalf of a named user with that user's permissions. Attribute-based rules make this practical at scale: a document filter can be

© 2026 floragunn GmbH · search-guard.com

built from the user's own attributes, such as department, region or clearance, taken from JWT claims, LDAP, SAML or OpenID Connect. Search Guard FLX 4.2.0, released in September, adds the option to fetch further claims from the OpenID Connect UserInfo endpoint when a token does not carry everything the rules need.

Encryption and Traceability for AI Data

Search Guard's Encryption at Rest works at the Lucene level and covers index files, snapshots and translogs. The translog matters for AI pipelines in particular, because it holds recently ingested content before it is committed. Encryption keys stay under customer control, with no external key service and no vendor-held escrow.

Audit logging can record which user or service accessed which index and documents, and when. The same records that support GDPR, ISO 27001 or SOX reporting in conventional deployments apply unchanged to AI workloads, which gives governance teams the traceability that emerging AI rules are starting to expect.

Relevant Capabilities

- Document- and field-level security, including attribute-based rules from JWT, LDAP, SAML and OpenID Connect
- Additional user claims from the OpenID Connect UserInfo endpoint (new in FLX 4.2.0)
- User impersonation, so an AI service can query with the permissions of the person asking
- Search Guard auth tokens and JWT authentication for services and integrations
- Encryption at Rest for index files, snapshots and translogs, with customer-held keys
- Audit logging of data access by users and services
- Licensing per production cluster, independent of node count, so growing AI indices on an existing cluster do not raise the licence cost

> "An AI assistant can only leak what its search layer hands it. Search Guard makes sure Elasticsearch hands it what the user may read, and logs that it did."
>
> **Hendrik Saly, CTO, floragunn GmbH**

Learn more: https://search-guard.com/security/

Attribute-based document-level security: https://docs.search-guard.com/latest/attribute-based-dls

User impersonation: https://docs.search-guard.com/latest/user-impersonation

FLX 4.2.0 changelog: https://docs.search-guard.com/latest/changelog-searchguard-flx-4.2.0

This article was published on the Search Guard blog: https://search-guard.com/blog/search-guard-secures-retrieval-layer-ai-applications-elasticsearch/

Search Guard is Security and Alerting for Elasticsearch and Kibana — search-guard.com · Start a free trial