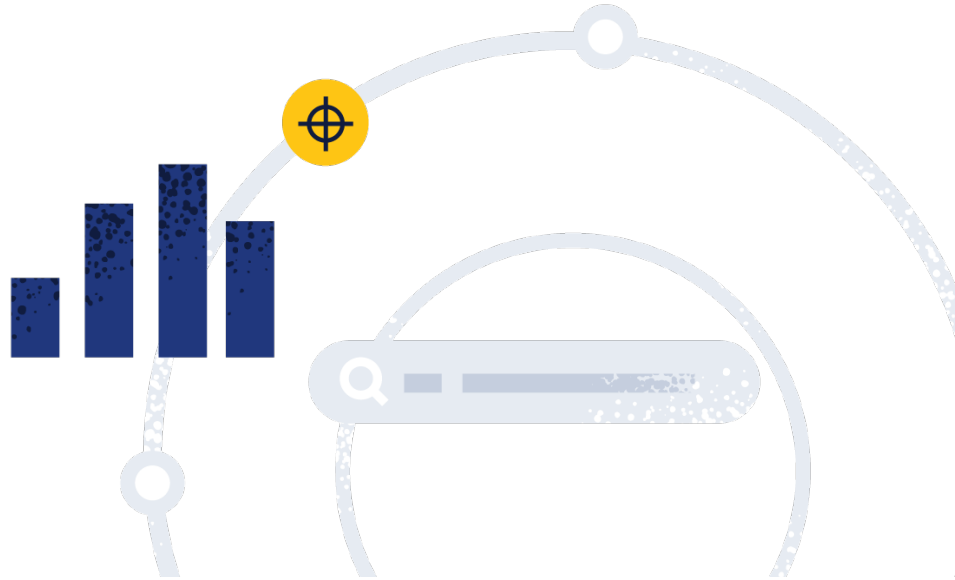
elastic

# KI-gestützte Sucherlebnisse erstellen

Eine Blaupause für Ihre erfolgreiche Suche

# Inhaltsverzeichnis



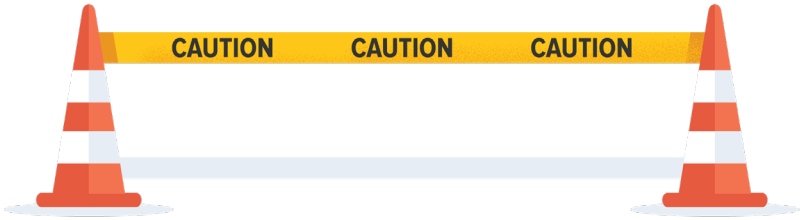


# Die Baustelle (für Ihre Suche)

Die Erstellung einer Suchfunktion für Ihre App oder Ihre Inhalte und der Bau eines Hauses haben vieles gemeinsam. Kommunikation, Planung und Vorbereitung sind unverzichtbar. Das Fundament ist entscheidend für eine langlebige und stabile Struktur. Architektur spielt eine wichtige Rolle. Das Geschehen hinter den Kulissen – beziehungsweise hinter den Wänden – sorgt dafür, dass alles reibungslos läuft: Wartung, Feinjustierung, Instandhaltung und Updates. Kleine, personalisierte Details können einen riesigen Unterschied machen.

Technologie definiert die Erwartungen von Hausbesitzern und Suchanwendern fortlaufend neu. Moderne Häuser mit effizienten, intelligenten Systemen und Materialien. Intelligente Sucherlebnisse, die anstelle von einfachen, schlüsselwortbasierten Abfragen KI-gestützte Relevanz und Antworten in Textform liefern und einem menschenähnlichen Verständnis immer näher kommen.




Letztendlich sind keine zwei Häuser oder Sucherlebnisse genau gleich. Insbesondere, nachdem sie eine Zeitlang „bewohnt“ und über diese Zeit hinweg gepflegt und personalisiert wurden, um sie einem echten Zuhause anzunähern.

Diese Anleitung ist eine Art von Blaupause und befasst sich damit, wie Sie im Zeitalter der KI ein solides Fundament für erfolgreiche Sucherlebnisse erstellen können.

# Suchrelevanz

Wenn mit dem Fundament etwas nicht stimmt, können alle möglichen strukturellen Integritätsprobleme für das darauf stehende Haus entstehen. Dasselbe gilt auch für die Suchrelevanz, die als zentrales Fundament für jedes solide gebaute Sucherlebnis dient.

Die Suchrelevanz misst, wie gut die Suchergebnisse den Absichten und Erwartungen der Nutzer entsprechen. Das Ziel besteht nicht nur darin, die richtigen Informationen zu finden, sondern auch, die Ergebnisse möglichst aussagekräftig zu ordnen und relevante Ausschnitte hervorzuheben. Systeme mit generativer KI können auch direkte, hilfreiche Antworten auf menschenähnliche Weise in Textform zurückgeben.

Relevanz ist nicht nur komplex, sondern in gewisser Weise auch eine Grauzone. Die Menschen haben sehr unterschiedliche Vorstellungen davon, was relevant ist, und formulieren ihre Fragen teils sehr unterschiedlich. Ohne zusätzlichen Kontext ist es schwierig, genau zu verstehen, nach welchen Antworten Ihre Nutzer suchen. Relevanz hängt auch mit Personalisierung zusammen: Wenn die Suchergebnisse die persönlichen Vorlieben der Nutzer berücksichtigen, ist das Ergebnis oft relevanter.

Herkömmliche Suchfunktionen gleichen Schlüsselwörter ab, um Ergebnisse für die Abfragen der Nutzer zu liefern. Die Nutzer müssen die genauen Schlüsselwörter kennen, um optimale Ergebnisse zu erhalten, und müssen die Rückgabe anschließend sortieren, um zum gewünschten Ergebnis zu gelangen.

KI ist jedoch dabei, die Art und Weise, wie wir Sucherlebnisse messen, zu transformieren.
Mit KI- und ML-Techniken (Machine Learning) können Sie exaktere Such-Tools erstellen, die Relevanz und Personalisierung verbessern, Abfragen in natürlicher Sprache und Nutzerabsichten besser verstehen und immer menschenähnlichere Antworten liefern.

# Vektorsuche

Die Vektorsuche ist ein fundamentales Feature von Vektordatenbanken und dient als Schlüsselkomponente für KI-gestützte Suchfunktionen.

Im Gegensatz zu herkömmlichen Suchmethoden mit Schlüsselwörtern, lexikalischer Ähnlichkeit und Worthäufigkeiten verwendet die Vektorsuche KI und ML, um die semantischen Beziehungen zwischen Wörtern und Dokumenten zu analysieren. Anstatt also Schlüsselwörter abzugleichen, erstellt die Vektorsuche eine mathematische Darstellung von Dokumenten und Abfragen, um Kontext, Relevanz und Ähnlichkeit von verschiedenen Informationsfragmenten besser zu verstehen.

Letztendlich können wir damit Dokumente oder Datenpunkte vergleichen, zwischen denen keine expliziten Verbindungen oder Beziehungen existieren. Mit den Zuordnungen der Vektorsuche können Sie beispielsweise Produkte finden (und empfehlen), die einem anderen Produkt ähneln, an dem ein Nutzer Interesse gezeigt hat.

Erweitern Sie das Vektorsucherlebnis zusätzlich, indem Sie die Vektorsuche mit dünnbesetzten Vektormodellen wie BM25 oder SPLADE kombinieren, um ein multimodales Sucherlebnis bereitzustellen, indem Sie die Ergebnisse nach Vektorähnlichkeit ordnen und gleichzeitig Texte besser klassifizieren können.

Mit der Vektorsuche können Sie auch große Sprachmodelle (Large Language Models, LLMs) und generative KI-Anwendungen ergänzen, da diese Systeme Fragen beantworten können, indem sie Dokumente in Texteinbettungen umwandeln und exakte Antworten liefern.

## Vektordatenbanken

Vektordatenbanken sind ideal, um unstrukturierte Daten zu verarbeiten und aussagekräftige Einblicke daraus zu extrahieren, und um sicherzustellen, dass Sie ein agiles Sucherlebnis entwickeln.

Eine Vektordatenbank ist eine Datenverwaltungslösung zum Speichern und Filtern von Metadaten und ist skalierbar, unterstütz dynamische Datenänderungen, führt Sicherungen durch und bietet integrierte Sicherheitsfunktionen.

Diese Datenbanken speichern und verwalten unstrukturierte Daten, wie etwa Text, Bilder oder Audio in Vektoreinbettungen, die man auch als hochdimensionale Vektoren bezeichnet und die von LLMs oder anderen KI-Modellen generiert werden. Diese auch als „dichte“ Vektoren bezeichneten Einbettungen sind numerische Darstellungen von Datenobjekten, die als Eingabe für ML-Algorithmen dienen, um die semantische Ähnlichkeit zu ermitteln.

Mit diesen Einbettungen können Vektordatenbanken riesige Mengen an unstrukturierten und teilweise strukturierten Daten (Daten, die keinem Datenmodell folgen, aber eine gewisse Struktur haben) indexieren und durchsuchen. Vektordatenbanken werden zur Verwaltung von Vektoreinbettungen erstellt und bieten daher eine vollständige Lösung zur Verwaltung unstrukturierter und teilweise strukturierter Daten.

All dies klingt zwar komplex, aber Vektordatenbanken wurden speziell für Entwickler erstellt und verwenden APIs, um eine nutzerfreundliche Schnittstelle bereitzustellen, die das Vektorsucherlebnis vereinfacht. Vektordatenbanken bieten noch zahlreiche weitere Vorteile:

- Skalierbarkeit für wachsende Datenvolumen
- Unterstützung für Echtzeit-Datenaktualisierungen für dynamische Änderungen an Daten
- Unterstützung für routinemäßige Sicherungen sämtlicher Daten in der Datenbank

## Bildähnlichkeitssuche

Ein beliebter Anwendungsfall besteht darin, ein Bild hochzuladen und ähnliche Bilder in einem Datensatz zu finden. Dazu wird der k-Nearest-Neighbor-Algorithmus (kNN) verwendet, ein beliebter Algorithmus, der die k nächstgelegenen Vektoren zu einem Abfragevektor findet.

Für große Datensätze, wie sie üblicherweise in Bildsuchanwendungen verarbeitet werden, benötigt kNN jedoch Unmengen an Rechenressourcen, und die Ausführung kann unverhältnismäßig lang dauern. Die Lösung ist der geschätzte nächste Nachbar (Approximate Nearest Neighbor, ANN). Dieser Algorithmus ist nicht zu 100 % genau, wird jedoch auch in hochdimensionalen Einbettungsräumen sehr effizient und skalierbar ausgeführt.

# Generative KI

Generative KI war bis vor Kurzem noch Science-Fiction und Theorie vorbehalten, hält inzwischen jedoch immer stärker Einzug in unseren Alltag. Der Hype um diese Technologie beschränkt sich zwar hauptsächlich auf Tools wie ChatGPT, DALL-E und Bard, aber nur wenige Anwendungsfälle für generative KI haben so viel Begeisterung hervorgerufen wie die generative Suche.

Um die Power der generativen KI nutzen zu können, benötigen Unternehmen ein speziell für ihre Umgebung justiertes Modell anstelle der Verbrauchermodelle, die such auf öffentlich verfügbare Trainingsdaten beschränken und keine bereichsspezifischen Daten, Sprachen und Inhalte kennen.

Die Nutzung von generativen KI-Modellen, die mit generischen und öffentlich verfügbaren Daten trainiert wurden, ist auch bedenklich im Hinblick auf Datensicherheit, Zugriffssteuerung, Datenschutz und potenzielle Verzerrungen, bei denen ein KI-Modell Inhalte auf Basis von Datensätzen produziert, die menschliche Tendenzen enthalten.

Für Entwicklungsteams erscheint die Erstellung generativer Sucherlebnisse aufgrund der Komplexität im Zusammenhang mit LLMs und generativer KI oft als unerreichbares Ziel. Damit sind Sie nicht allein.

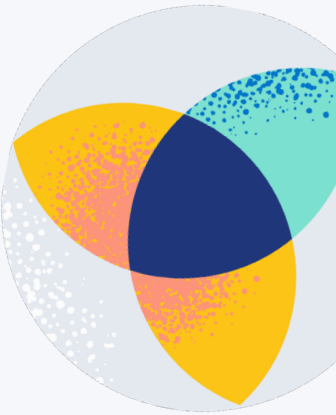
 

## Mögliche Herangehensweisen:

**Erstellen und Trainieren eines bereichsspezifischen Modells von Grund auf**
Diese Herangehensweise erfordert jedoch umfangreiche finanzielle Ressourcen, Zugang zu beträchtlicher Rechenleistung auf spezialisierten Chips sowie KI-Spezialkenntnisse. Auch wenn Sie über das Budget und die KI-Kenntnisse verfügen, ist es immer noch schwierig, große Mengen an hochwertigen Trainingsdaten für ein LLM zu finden.

**„Feinjustierungs"-Training eines vorhandenen LLM**
Bei diesem Ansatz fügen Sie bereichsspezifische Inhalte zu einem System hinzu, das bereits mit allgemeinem Wissen und sprachbasierten Interaktionen trainiert wurde. Dazu sind typischerweise weniger Daten – Hunderte oder Tausende von Dokumenten anstelle von Millionen oder Milliarden – und weniger Rechenzeit erforderlich als zur Erstellung eines völlig neuen Modells. Dabei sind jedoch auch Einschränkungen zu berücksichtigen. Obwohl dieses Modell weniger Rechenleistung benötigt, kann der Vorgang dennoch recht teuer sein. Außerdem erlauben manche LLM-Anbieter (zum Beispiel OpenAI) keine Feinjustierung ihrer neuesten LLMs, wie etwa GPT-4.

**Textgenerierung mit proprietären Datenquellen ergänzen**
Um die Herausforderungen der generativen Suche zu überwinden, können Unternehmen generative KI-Modelle mit ihren proprietären Daten verbinden, um die kontextbezogene Genauigkeit der Ergebnisse zu verbessern.



Auf diese Weise können Sie die Einschränkungen herkömmlicher Suchfunktionen überwinden und Ihre Daten effektiver und umfangreicher nutzen. Sie können die Daten in Unterhaltungs-Apps anwenden, die komplexe Fragen beantworten, exakte Zusammenfassungen aus vielen verschiedenen Quellen liefern und den Nutzern die benötigten Informationen schneller liefern.

# Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG) wird immer häufiger als Lösung für viele der Probleme (und die hohen Kosten) im Bereich der generativen KI eingesetzt. Diese Technik ist eine Kombination aus LLMs und Informationsabruftechnologien und ermöglicht KI-Interaktionen unter Berücksichtigung der proprietären Daten Ihres Unternehmens.

Kurz gesagt überbrückt das KI-Framework die Lücke zwischen den LLMs, auf denen die generative KI basiert, und den privaten Datenquellen. RAG verwendet ein Abrufmodell, das relevante Informationen in vorhandenen privaten oder proprietären Datenquellen findet, und das generative Modell nimmt die abgerufenen Informationen entgegen, synthetisiert sämtliche Daten und bildet daraus eine zusammenhängende und dem Kontext angemessene Antwort.

RAG bietet außerdem leistungsstarke semantische Fähigkeiten, um kontextbezogene Daten in eine Abfrage einzufügen, bevor sie an ein LLM übermittelt wird. Das System speichert bereichsspezifisches Wissen, um hochwertige Ergebnisse zu liefern, und Sie können herkömmliche Suchergebnisse mit generativer KI kombinieren, um Fragen zu beantworten.

## Ein Framework für die KI-gestützte Suche

### Private Daten

- Ein Bestand an privaten oder proprietären Informationen, wie etwa interne Unternehmensdokumente, Richtlinien oder Geschäftstransaktionen

### Abruftechnologie mit RAG

- Schlüsselwortsuche, Vektorsuche oder eine Kombination aus beidem, um relevante Informationen in vorhandenen privaten Informationsquellen zu finden
- Nutzung semantischer Fähigkeiten, um kontextbezogene Daten in eine Abfrage einzufügen, bevor sie an ein LLM übermittelt wird
- Erweiterung der Ausgabe des Sprachmodells, ohne das Modell neu zu trainieren

### LLM für generative KI

- Synthetisiert abgerufene Informationen in eine dem Kontext angemessene Antwort
- Reagiert auf Anfragen mit semantisch relevanten Antworten, oft unter Berücksichtigung von vorherigen Anfragen und Kontext

# Fallstudie: CISCO

**Technologien von Cisco Systems sind bei mehr als 87 % aller Fortune 500 -Unternehmen im Einsatz. Mit der Website-Suche von Cisco finden die Besucher Informationen in hunderttausenden Webseiten, Dokumenten sowie technischen, produkt- und unternehmensspezifischen Ressourcen. Genauigkeit und Relevanz sind für diese Suche entscheidend.**

## Vorher

11.000 Supportmitarbeiter benötigten Such-Tools, um Inhalte aus Millionen von Dokumenten abzurufen und mehr als zwei Millionen Serviceanfragen pro Jahr zu bearbeiten. Genauigkeit und Geschwindigkeit der zurückgegebenen Suchergebnisse sind entscheidend. Eine Verzögerung von nur einer halben Sekunde kann sich bereits negativ auf die Click-Through-Rate der Website oder das Kundenerlebnis beim Kontakt mit einem Supportmitarbeiter auswirken.

## Nachher

Mit einer neuen, KI-gestützten Architektur für die Unternehmenssuche hat Cisco ein völlig neues Kundenserviceerlebnis entwickelt. Mit diesem Such-Tool können die Supportmitarbeiter im Handumdrehen die verfügbare Dokumentation für eine Kundenserviceanfrage abrufen und pro Monat bis zu 5.000 Arbeitsstunden einsparen. Mit der KI-gestützten Suche konnte Cisco auch die Website-Suche des Unternehmens für Endnutzer überarbeiten. Wenn Nutzer jetzt eine Suchanfrage eingeben, wird eine Dropdown-Liste mit automatischen Vorschlägen angezeigt und in Echtzeit aktualisiert, wenn weitere Zeichen und Wörter eingegeben werden. Dazu gehören auch allgemeine Fragen, die die Suchabsicht wiedergeben und mit denen die Suchenden relevante Informationen schneller finden können.

**73 %** schnellere Antwortzeiten auf Suchabfragen

**Vollständige Story lesen**

## TL;DR: Vorteile von Retrieval Augmented Generation (RAG)

- **Einheitlicher Kontext:** Der Abrufprozess garantiert, dass die generierte Ausgabe kontextrelevant und für die entsprechende Eingabe und den abgerufenen Kontext einheitlich ist.
- **Bessere Genauigkeit:** RAG ruft relevante Informationen zuerst ab und liefert daher genauere Antworten, insbesondere beim Informationsabruf und bei der Beantwortung von Fragen.
- **Steuerbarkeit:** Mit dem Abrufmechanismus können Sie den Umfang der generierten Antworten steuern, indem Sie die entsprechende Abrufdatenbank auswählen oder bestimmte Abfragen vorgeben.
- **Attraktivität:** RAG-basierte Sucherlebnisse sind für die Nutzer sehr attraktiv. Mit Details und Kontext aus internen privaten Daten angereicherte RAG-basierte Systeme liefern aufschlussreiche und ansprechende Antworten.
- **Weniger Verzerrung in den Ergebnissen:** Mit dem abgerufenen Kontext kann RAG einige der Verzerrungsprobleme minimieren, die in rein generativen Modellen häufig auftreten.

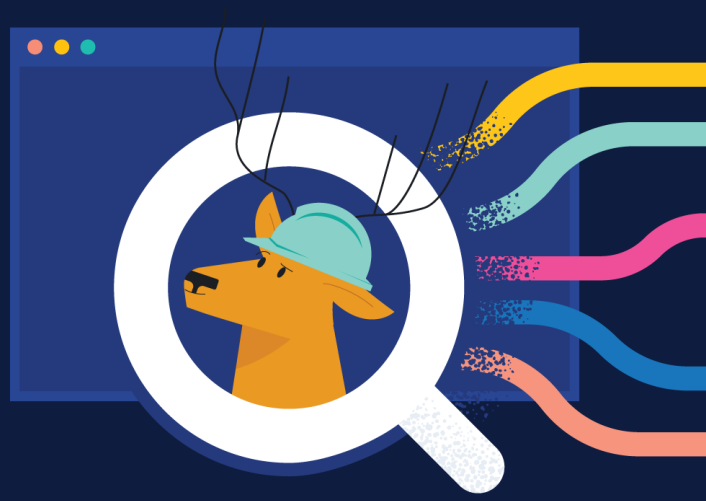
 

# Daten und Dateningestion

Die Sanitärinstallation in Ihrem Haus ist ein Komfort, auf den Sie kaum verzichten können. Einfacher Zugang zu Trinkwasser, Toiletten und eine Waschmaschine sind nicht nur wichtig, sondern zum Teil sogar lebensnotwendig. Mit den Eingangsdaten für Ihr Sucherlebnis verhält es sich ähnlich.

Diese Daten sind möglicherweise chaotisch, unvollständig, in Silos verteilt oder auf andere Arten unzugänglich, was das Erlebnis für Ihre Nutzer beeinträchtigt. Um dies zu beheben, können Sie die Daten, in denen Ihre Nutzer später suchen, mit einer Ingestions-Pipeline (wie in der Sanitärtechnik) ingestieren, anreichern, transformieren und synchronisieren. Ohne diese unverzichtbaren Dateningestionsmechanismen können Sie kein ansprechendes Sucherlebnis bereitstellen. Mit APIs, Web-Crawler, Connectoren und Pipelines haben Sie alle benötigten Tools um die zu durchsuchenden Daten sammeln und integrieren zu können.

Bevor wir uns näher mit den Ingestions-Tools befassen, werfen wir einen genaueren Blick auf die Arten von Daten, denen Sie begegnen werden.



# Strukturierte Daten

Ein Großteil der Unternehmensdaten liegt als strukturierte Daten vor. Diese Daten sind gut organisiert und formatiert und können daher von Machine-Learning-Algorithmen und von Menschen zügig verarbeitet werden. Beispiele für strukturierte Daten sind Metriken, Datumsangaben, Namen, Postleitzahlen und Kreditkartennummern.

## Gute Zeichen

- **Hilft, das Suchverhalten der Kunden** anhand von Datenpunkten wie Namen, Kaufhistorie und Aufenthaltsort auszuwerten
- **Unterstützt das Kundenbeziehungsmanagement (Customer Relationship Management, CRM)**, bei dem Unternehmen das Verhalten ihrer Kunden mit relationalen Datenbanken analysieren
- **Einfache Speicherung** in relationalen Datenbanken, NoSQL-Datenbanken, Data Warehouses, Data Lakes, speicherinternen Datenbanken und anderen Formaten dank der vorhandenen Struktur
- **Belegt weniger Platz** als unstrukturierte Daten
- **Einfacheres Crawling für Machine-Learning**-Algorithmen
- **Liefert oft bessere** und genauere Business Intelligence aufgrund der einfachen Analysierbarkeit

## Warnzeichen

- **Eingeschränkt** und nur für den vorhergesehenen Zweck nutzbar
- **Oft von minderwertiger Qualität** aufgrund der riesigen Datenmengen in Unternehmen. Außerdem entstehen oft Duplikate oder Daten, die nicht mehr relevant sind.

# Unstrukturierte Daten

Unstrukturierte Daten existieren in den meisten Unternehmen im Überfluss – und werden ständig neu generiert.

**80 %**

**80 % aller vorhandenen Daten sind unstrukturiert.***

**430 %**

**Unstrukturierte Daten werden zwischen 2018 und 2025 um 430 % anwachsen.***

Wie der Name bereits verrät, folgen unstrukturierte Daten keinem herkömmlichen Datenmodell und werden typischerweise in ihrem nativen Format gespeichert. Unstrukturierte Daten sind meistens als qualitativ kategorisiert und können von Menschen oder von Maschinen generiert werden.

Beispiele für unstrukturierte Daten sind Informationen aus IoT-Geräten (Internet der Dinge), wie etwa Ticker- oder Sensordaten, Textdaten, wie etwa E-Mails oder Rechnungen, wissenschaftliche Daten, wie etwa computergenerierte Weltraumerkundungsdaten oder seismische Berichte, sowie Daten und Bilder aus dem Gesundheitswesen, wie etwa Kernspin-, Röntgen- oder CT-Scans.

*Quelle: Unstructured Data Storage [UDS] Survey, IDC, 2021

# Unstrukturierte Daten

## Gute Zeichen

- **Enthält oft umfassende und ausführliche Informationen,** die in strukturierten Daten nicht verfügbar sind
- **Hilft Unternehmen, ihre Kunden** und Marktverschiebungen besser zu verstehen, um ihre NLP-Modelle (Natural Language Processing, NLP) zu trainieren und prädiktive Datenanalysen zu liefern. E-Commerce-Unternehmen können beispielsweise Kundeninteraktionen nachverfolgen und Echtzeitdaten sammeln, um Ausgabenmuster zu identifizieren, personalisierte Erlebnisse zu erstellen und Preisstrategien zu entwickeln.
- Liefert Unternehmen **tiefere Einblicke** in die Absichten und Verhaltensweisen der Kunden

## Warnzeichen

- **Erfordert Analysen,** um wertvolle Erkenntnisse zu liefern
- **Unorganisiert** und umfangreich, was die Analyse erschwert
- **Schwer, manuell zu verwalten**
- Viele Datenbanken und Tools **sind dem Volumen und der Vielfalt nicht gewachsen**.
- **Spezielle Tools und Technologien sind erforderlich**, um das exponentiell wachsende Datenvolumen zu bändigen.
- **Die Qualität ist oft uneinheitlich.** Enthalten oft Fehler, Unstimmigkeiten oder irrelevante Informationen enthalten, was es erschwert, genaue Informationen zu extrahieren, insbesondere für Suchanwendungen.
- **Das Vorverarbeiten oder Bereinigen unstrukturierter Daten** zur Verbesserung der Qualität ist oft zeitraubend und komplex.

## Warum wird die Schlüsselwortsuche immer noch verwendet, wenn RAG und KI-Suche so gut sind?

Die schlüsselwortbasierte oder lexikalische Suche existiert seit Jahrzehnten und ist nach wie vor eine wichtige Komponente für Sucherlebnisse. Auch wenn sich moderne, KI-gestützte Techniken immer weiter ausbreiten.

Die lexikalische Suche eignet sich immer noch hervorragend, um exakte Übereinstimmungen zu finden, etwa für Produkt-SKUs, Fehlercodes, Nutzer-IDs, Supporttickets, Codeausschnitte und so weiter.

Ein weiterer Grund für die anhaltende Beliebtheit der Schlüsselwortsuche ist die große Anzahl an Suchprodukten und -Apps, die diese Technologie nach wie vor verwenden. Es ist nicht einfach und oft auch nicht notwendig, neue Ansätze wie die Vektorsuche oder die semantische Suche für diese Anwendungen einzusetzen.

Mit einer Kombination aus Schlüsselwort-, Vektor- und semantischer Suche erhalten Sie das Beste aus beiden Welten: exakte Übereinstimmungsrelevanz zusammen mit natürlichem Sprachverständnis, um die Nutzerabsichten zu verstehen. **Dieser Ansatz ist oft notwendig, um neuere Daten mit vorhandenen Systemen und Daten zu kombinieren.**

SPICKZETTEL

# Verwalten und Analysieren von unstrukturierten Daten

Unstrukturierte Daten haben naturgemäß keine vordefinierte Struktur, die das Verwalten und Analysieren der Daten vereinfachen könnte. Um diese Daten analysieren zu können, müssen Sie also zunächst eine Struktur für deren Verwaltung definieren, um sie anschließend speichern, organisieren und sichern zu können.

Sie haben eine Vielzahl an Speichertechnologien und Datenverarbeitungsmethoden zur Auswahl, mit denen Sie unstrukturierte Daten speichern, verwalten und analysieren können.

## Datenverarbeitungsmethoden

- **Natürliche Sprachverarbeitung (Natural Language Processing, NLP)**
  Diese Technologie befasst sich mit der Interaktion zwischen Computern und Menschen in Form von natürlicher Sprache. NLP hat das Ziel, menschliche Sprache möglichst sinnvoll lesen, entziffern, verstehen und interpretieren zu können.
- **Machine Learning (ML)**
  ML verwendet statistische Methoden, um Muster in strukturierten und unstrukturierten Daten zu erkennen und Vorhersagen oder Entscheidungen zu treffen.

## Datenspeicherungstechnologien

- **Data Lakes**
  Aufgrund der Vielfalt und des großen Volumens können unstrukturierte Daten in Data Lakes oder direkt am Ort der Datenerstellung (Edge) gespeichert werden. Data Lakes eignen sich hervorragend, um große Mengen verschiedener Datentypen zu speichern.
- **Content-Management-Systeme (CMS)**
  Mit einem CMS können Unternehmen unstrukturierte Daten speichern, abrufen, durchsuchen, indexieren und online veröffentlichen.

# Ingestion

Nachdem Sie entschieden haben, welche Daten Sie durchsuchen möchten, können Sie sie im nächsten Schritt in Ihre Suchmaschine integrieren. Bei der Dateningestion werden große Mengen geordneter Daten von verschiedenen Orten – oder aus Ihren Informationsquellen – erfasst. Dieser Prozess ist oft mühsam. Durch die zunehmende Menge und Vielfalt der Datenquellen ist das Sammeln, Zusammenstellen und Transformieren von Daten in eine zusammenhängende und brauchbare Form eine ständige Herausforderung. Zum Beispiel:

- **Datenintegration über verschiedene Quellen hinweg:** Um Informationen aus verschiedenen Systemen des Unternehmens zu integrieren, müssen unterschiedliche Datenanforderungen und Standards miteinander synchronisiert werden, was oft chaotisch abläuft. Die Daten können in unterschiedlichen Formaten und an unterschiedlichen Orten vorliegen, wie etwa relationale Datenbanken, Webseiten, Dateisysteme oder Netzlaufwerke.
- **Datenschutzbedenken:** Möglicherweise existieren Metadaten, die Sie den Endnutzern auf keinen Fall zeigen möchten, oder personenbezogene Informationen, die vor dem Ingestieren bereinigt werden müssen.
- **Fehlerhafte Daten:** Oft enthalten die Daten auch Fehler, Unstimmigkeiten oder fehlende Werte, die vor dem Ingestieren identifiziert und korrigiert werden müssen.
- **Geschwindigkeit und Aktualität der Daten:** Wie viele tage oder Monate an Geschäftseinträgen möchten Sie in der Ebene vorhalten, die eine schnelle Suche ermöglicht? Wann sollen die Daten in eine Archivebene übertragen werden, die zwar langsamer, dafür aber kostengünstiger ist?
- **Manuelle Ingestion:** Es kann sehr mühselig sein, den Code zum Ingestieren der Daten zu schreiben und Mappings zum Extrahieren, Bereinigen und Laden der Daten manuell zu erstellen.
- **Kosten:** Die zur Unterstützung verschiedener Datenquellen und Tools erforderliche Infrastruktur kann auf lange Sicht hohe Kosten verursachen.

Je nach Datenquelle haben Sie eine Vielzahl von Mechanismen für die Dateningestion zur Auswahl. Werfen wir einen genaueren Blick auf die Optionen.

**Web-Crawler**

Beim Crawling werden Daten aus einer oder mehreren Datenquellen importiert, geladen und in einem strukturierten Format bereitgestellt, um sie mit einer Suchmaschine indexieren und durchsuchen zu können. Bei diesem Verfahren werden die Daten auch wiederholt abgerufen (in Echtzeit oder als Batch-Vorgang), um die Aktualität der Suchergebnisse sicherzustellen. Ein Web-Crawler ist ein digitaler Suchmaschinen-Bot, der Text- und Metadaten verwendet, um Seiten auf Websites zu entdecken und zu indexieren. Er durchforstet eine angegebene Domain, um den Inhalt einer Seite zu ermitteln. Anschließend werden die Seiten indexiert und die Informationen zur späteren Verwendung gespeichert.

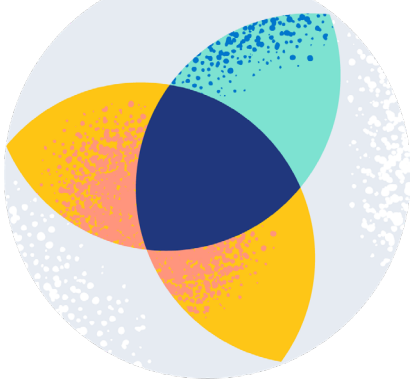
**API**

Nutzen Sie API-Endpoints, um einen programmgesteuerten Ingestionsablauf zwischen Ihrer Datenquelle und der App zu erstellen, die die Daten verwendet. Mit den standardisierten CRUD-Operationen können Sie die Daten nach Belieben hinzufügen, aktualisieren und entfernen. Nutzen Sie Client-Bibliotheken sowie Such-, Sortier- und Filterfunktionen aus den nativen APIs der Programmiersprachen, um den Prozess zu vereinfachen.

**Connectors**

Connector-Bibliotheken synchronisieren Daten aus einer ursprünglichen Datenquelle in Ihren benötigten Index, zum Beispiel ein Netzlaufwerk oder eine PostgreSQL-Datenbank. Connectors extrahieren die ursprünglichen Dateien, Einträge oder Objekte und transformieren sie in die benötigten Dokumente.



**Ingestionspipelines**

Mit Pipelines können Sie gängige Transformationen auf Ihre Daten anwenden, bevor Sie sie speichern. Sämtliche Ingestionspipelines bestehen aus einer geordneten Liste an Prozessoren, die das Verhalten der Pipeline definieren. Dies ist oft sehr hilfreich, um eine zusätzlich Anpassungs- und Nachbereitungsebene für Dokumente bereitzustellen. Zum Beispiel:

- Einheitliche Extraktion von Texten aus binären Datentypen
- Einheitliche Formatierung
- Einheitliche Bereinigungsschritte (Entfernen personenbezogener Daten, wie etwa Telefonnummern oder Sozialversicherungsnummer)

Pipelines enthalten auch immer öfter ML-Prozessoren, um beim Ingestieren der Daten Rückschlüsse zu ziehen. Inferenzaufgaben können beispielsweise Stimmungen oder Absichten aus einem Datensatz mit Nutzerkritiken ermitteln.

Es ist oft sehr mühsam, produktionsreife Pipelines von Grund auf einzurichten und zu verwalten. Faktoren wie Fehlerbehandlung, bedingte Ausführung, Abfolge, Versionierung und Modularisierung müssen berücksichtigt werden.

# Personalisierte Nutzererlebnisse

Der Ort, an dem Sie leben, ist mehr als nur eine Struktur, eine Wohnung oder ein Gebäude. Er ist Ihr Zuhause. Durch die Personalisierung dieser Bereiche entstehen Komfort und Gemütlichkeit. Möglicherweise nutzen Sie Feng-Shui, um Ihr Zuhause harmonisch einzurichten. Vielleicht bevorzugen Sie auch eine stimmungsvolle Beleuchtung, einen gemütlichen Kamin oder Ihre Lieblingskunstwerke. Mit dem Nutzererlebnis in Suchanwendungen verhält es sich ähnlich. Ihre Kunden, potenziellen Kunden und Nutzer sollen sich wie zu Hause fühlen mit vertrauten, nützlichen, schnellen, leitenden und relevanten Suchfunktionen. Helfen Sie ihnen, die gesuchten Informationen zu finden – und vielleicht auch einige Dinge, von denen sie nicht einmal wussten, dass sie sie brauchten. Die Nutzer bemerken es definitiv, wenn sich ein Erlebnis plump anfühlt, weniger relevante Ergebnisse liefert oder natürliche Sprache nicht richtig versteht.

# Semantische Suche

Dank semantischer Suche und LLMs sind moderne Suchmaschinen immer intelligenter geworden und liefern inzwischen höchst relevante und personalisierte Suchergebnisse.

Die semantische Suche wendet Nutzerabsicht und Bedeutung – auch als Semantik bezeichnet – auf Wörter und Sätze an, um passende Inhalte zu finden. Sie verwendet mehr als nur Schlüsselwörter, um zu verstehen, wonach die Nutzer tatsächlich suchen – anhand verschiedener Faktoren wie Standort, Suchverlauf und Trends –, um möglichst relevante Ergebnisse zurückzuliefern.

So entsteht ein personalisiertes Erlebnis, das die Suchergebnisse anhand der vorherigen Suchanfragen oder dem Aufenthaltsort der Nutzer ordnet. Wenn Sie beispielsweise „Restaurants“ in Google eingeben, werden gut bewertete Ergebnisse in Ihrer Nähe oder aus Kategorien angezeigt, nach denen Sie zuvor gesucht haben.

Diese Art der Personalisierung ist für uns inzwischen selbstverständlich. Natürlich meinen wir „Restaurants in meiner Nähe“, wenn wir nach „Restaurants“ suchen. Auf diese Weise sind unsere allgemeinen Erwartungen an Sucherlebnisse entstanden. Dennoch ist es bei der Erstellung semantischer Sucherlebnisse oft schwierig, einen Einstiegspunkt zu finden.

Unsere Empfehlung: Ermitteln Sie zunächst, ob Ihr Unternehmen über das nötige Datenvolumen verfügt, um ein extrem personalisiertes Erlebnis zu unterstützen. Sie benötigen Zugriff auf Daten, und zwar eine Menge davon. Dies ist die einzige Situation, in der weniger nicht mehr ist.

## Spezial-Tipp:

Achten Sie auch darauf, Erwartungen angemessen intern zu lenken. In vielen Unternehmen wird die Erstellung von Sucherlebnissen durch einen Mangel an Datenvolumen ausgebremst, wodurch wiederum die mögliche Anzahl an Dimensionen für Suchabfragen sinkt. Bei der Implementierung KI-gestützter Features sollten Sie Ihren Boss und die Geschäftsleitung auch daran erinnern, dass ein Machine Learning-Modell nur so gut ist, wie die Daten, mit denen es trainiert wurde.

# Fallstudie: LABELBOX

**Als Vorreiter der KI-Revolution bietet Labelbox eine kollaborative Datentrainingsplattform an, die beschriftete Daten für ML-Anwendungen erstellt und verwaltet. Anstatt eigene, teure Tools zu entwickeln, um Trainingsdaten zu erstellen oder zu verwalten, können die Kunden Labelbox verwenden.**

**Damit wird das Problem gelöst, KI- und ML-Initiativen aus der Forschungs- und Entwicklungsphase in die Produktion zu überführen. Neben der direkten Zusammenarbeit mit den Kunden erleichtert das Hauptprodukt des Unternehmens auch die Erstellung und Verwaltung beschrifteter Daten, um die Bereitstellung von KI-Anwendungen zu beschleunigen.**

## Vorher

Labelbox Catalog, eines der beliebtesten Tools des Unternehmens, durchsucht und analysiert unstrukturierte Daten, um die Performance von Trainingsmodellen zu verbessern. Catalog wurde jahrelang zusammen mit einer PostgreSQL-Datenbank ausgeführt, was die Erstellung von Filtern für exaktere Ergebnisse erschwerte. Suchvorgänge waren recht langsam: Manchmal dauerte es bis zu 20 Sekunden, eine Anfrage zu beantworten.

## Nachher

Die Kunden können jetzt komplexe Suchvorgänge ausführen, und ausgeklügelte Analysen unterstützen genauere Entscheidungen beim Auswählen der zu beschriftenden Daten. Die Suchdauer wurde auf etwa eine Sekunde reduziert.

**95 %** Schnellere Antwort auf Suchanfragen

**Vollständige Story lesen**

Durch die hohen Ansprüche im Hinblick auf die Personalisierung erwarten die Nutzer auch erweiterte Suchfunktionen als Standard. Dazu gehören Add-On wie Auto-Vervollständigung sowie die Möglichkeit, Ergebnisse zu filtern und zu sortieren, um genau das zu finden, wonach sie gesucht haben.

Filter sind inzwischen beispielsweise so allgegenwärtig, dass die Nutzer sie in praktisch jedem Sucherlebnis erwarten. E-Commerce-Kunden möchten auswählen können, wer den Artikel verkauft und in welcher Farbe und Größe er geliefert wird.

Die Implementierung von Filtern, mit denen die Nutzer ihre Suchergebnisse anhand verschiedener Attribute – Kategorien, Zeitpunkte, Preise usw. – eingrenzen können, erfordert jedoch sorgfältige Planung und Entwicklung.

Such-APIs sind der einfachste Weg, um moderne Suchfunktionen in Ihre Website oder Ihre Anwendung zu integrieren. Diese APIs unterstützen benutzerdefinierte Suchoberflächen und konfigurierbare Komponenten mit nur wenigen Codezeilen, wie etwa umfangreiche Filterfunktionen, „Search-as-you-type“ sowie automatische Paginierung. Sie unterstützen oft auch Clients in unterschiedlichen Sprachen, um die REST-APIs in praktisch jeder Programmiersprache Ihrer Wahl aufrufen zu können.

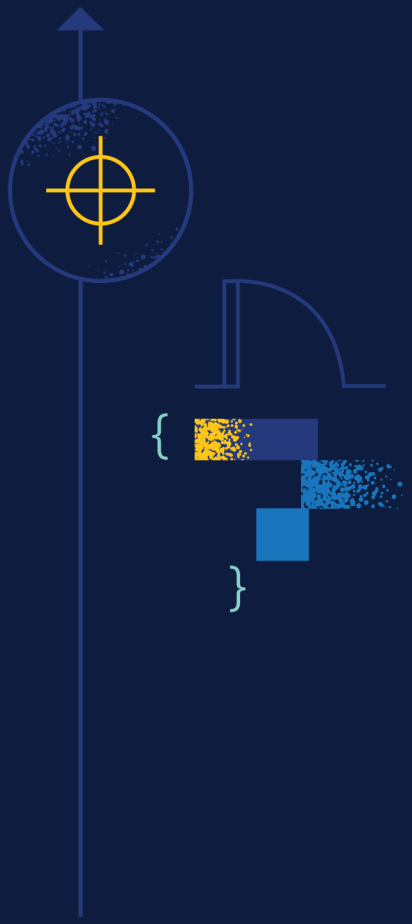
SPICKZETTEL

# Best Practices für UX in Suchanwendungen

Für die Nutzer ist es ein gravierender Mangel, wenn eines dieser grundlegenden Features in Ihrem Sucherlebnis fehlt:

- **Fehlertoleranz:** Ihre Suche muss eventuelle Rechtschreibschwächen oder Tippfehler berücksichtigen (insbesondere auf Mobilgeräten).
- **Auto-Vervollständigung:** Während der Eingabe einer Abfrage müssen entsprechende Vorschläge zur Auswahl angeboten werden.
- **Synonyme:** Machen Sie es Ihren Nutzern einfach, relevante Ergebnisse zu finden, egal ob sie nach Berg, Gipfel oder Gebirge suchen.
- **Filter/Facettensuche:** Bieten Sie Ihren Nutzern die Möglichkeit, Ergebnisse zu filtern und zu sortieren, um genau das zu finden, wonach sie gesucht haben.
- **Kuratierungen:** Sie müssen bestimmte Inhalte und Ergebnisse anheften, ausblenden oder hervorheben können, um die Nutzer zu leiten.
- **Paginierung oder Endlos-Scrolling:** Bieten Sie Ihren Nutzern Paginierung oder die Option, in den angezeigten Ergebnissen endlos zu scrollen.
- **Weitere Empfehlungen:** Insbesondere in E-Commerce-Anwendungen sind zusätzliche Vorschläge oder Empfehlungen sehr hilfreich, um zusätzliche Umsatzmöglichkeiten zu schaffen.

# Datenschutz und Zugriffssteuerung



Die Parallelen zwischen Datensicherheit und häuslicher Sicherheit sind ziemlich offensichtlich. Wir möchten uns in unserem Zuhause sicher fühlen und unsere Besitztümer schützen. Ebenso erwarten die Nutzer Ihrer Suchanwendungen, dass ihre Daten geschützt werden und dass standardmäßige Datenschutzprotokolle befolgt werden. Außerdem möchten Sie sicherstellen, dass die privaten Daten Ihres Unternehmens ebenfalls gut geschützt sind.

Dadurch entsteht eine zusätzliche Komplexitätsebene für Entwickler von agilen Sucherlebnissen. Die generative KI ist dabei, die Suche für Unternehmen zu revolutionieren, bringt jedoch auch eine Reihe von Herausforderungen im Hinblick auf Datenschutz und Sicherheit mit sich, da sie in der Lage ist, personenbezogene Informationen zu verarbeiten und potenziell vertrauliche Informationen zu generieren. Daher müssen Unternehmen und Entwickler, die mit KI-gestützten Sucherlebnissen experimentieren, auf den Schutz vertraulicher Daten achten, um potenzielle Katastrophen zu vermeiden.

Nehmen Sie beispielsweise an, dass die verwendeten Trainingsdaten vertrauliche Informationen wie Krankenakten, Finanzdaten oder sonstige Bezeichner enthalten. In diesem Fall besteht das Risiko, dass potenziell vertrauliche Informationen generiert werden, die gesetzliche Datenschutzbestimmungen verletzen und Einzelpersonen gefährden können. Ohne angemessene Sicherheitsvorkehrungen sind generative KI-Tools außerdem anfällig für Datenpannen, was zu unbefugten Zugriffen oder der Offenlegung vertraulicher Nutzerdaten und damit zu Datenschutzverletzungen und potenziellem Missbrauch personenbezogener Daten führen kann.

Entwickler müssen auch auf das Gleichgewicht zwischen Innovation und Datenschutz achten, um dem Missbrauch vertraulicher Informationen vorzubeugen. Dazu gehört die Integration von Systemen für Anomalieerkennung und Monitoring sowie die Durchführung fortlaufender Sicherheits-Audits, um die Einhaltung relevanter Datenschutzbestimmungen sicherzustellen.

Berechtigungen auf Dokumentebene sind ein weiterer wichtiger Punkt, den es bei der Erstellung unternehmensinterner Sucherlebnisse zu berücksichtigen gilt. Mit diesen Berechtigungen lassen Inhaltsfragmente anhand der Attribute von Einzelpersonen oder Teams verwalten, wie etwa private Dokumente der Personalabteilung, Beschaffungsverträge usw. In Kombination mit einer rollenbasierten Zugriffssteuerung (RBAC) bieten Berechtigungen auf Dokument- und Feldebene sehr effektive Mechanismen zur Verwaltung von Datenzugriffen.

Für die Implementierung dieser Features und die Durchführung verschiedener Datenschutzbeurteilungen sind nicht nur die Entwickler verantwortlich. Sie erfordern vielmehr die Zusammenarbeit zwischen IT-Sicherheit, IT-Wartung, IT-Systemmanagern und Analysten im gesamten IT-Bereich. Drittanbieter können auch beim Skalieren helfen und sicherstellen, dass Ihr Unternehmen die ständig wachsende Liste der Datenschutzbestimmungen automatisch einhält.

# Monitoring und Analyse

Ihr Zuhause kann aus wundervollen, hervorragend gestalteten und schön eingerichteten Räumen bestehen. Aber ohne funktionierende Heizung, Kühlung, Sanitäranlagen und Strom sitzen Sie im Dunkeln, es ist zu heiß oder zu kalt, Sie haben kein fließendes Wasser und die Lebensmittel im Kühlschrank sind verdorben.

Mit Suchanwendungen verhält es sich ähnlich. Suchrelevanz und Nutzeraktivitäten müssen regelmäßig überwacht und analysiert werden, um sicherzustellen, dass die Nutzer das finden, wonach sie suchen.

# Analysieren des Suchverhaltens

Wer etwas auf Ihrer Website sucht, hinterlässt dabei einen Schatz an verwertbaren Daten, die umfangreiche Möglichkeiten zur Suchanalyse bieten. Die Suchanalyse fördert zutage, wonach die Nutzer auf Ihrer Website suchen, auf welche Ergebnisse sie klicken, wie viel Zeit sie auf einer Seite verbringen und wie sich diese Journeys in Conversions auszahlen.

Die Analyse der Website-Suche liefert Einblicke wie etwa Suchabfragen ohne jegliche Ergebnisse, um Ihnen zu zeigen, wo es Lücken zu schließen gilt. Möglicherweise suchen Ihre Besucher nach Themen, über die Sie sich noch keine Gedanken gemacht haben, oder nach Produkten, die Sie noch nicht anbieten oder herstellen. Diese Art von Informationen können Sie anschließend an verschiedene Beteiligte in den Marketing- oder Produktentwicklungsabteilungen weitergeben.

Für Entwickler empfehlen wir eine umfangreiche, vorkonfigurierte Lösung, die Funktionen für Datenerfassung, Metriken und Visualisierungen des Suchverhaltens der Nutzer enthält. Auf diese Weise können Sie automatisch umfangreiche Analysen erfassen und diese Ergebnisse in einem benutzerfreundlichen Dashboard analysieren.

## Technisches Detail: Messen der Effektivität Ihrer Website-Suche

- **Click-Through-Rate (CTR):** Im Hinblick auf die Website-Suche verrät uns die CTR das Verhältnis zwischen den Nutzern, die auf ein Suchergebnis geklickt haben, und den Nutzern, die sich die Suchergebnisse angesehen haben. Mit diesem Verhältnis können Sie herausfinden, ob die Nutzer bei ihrer Suche interessante und relevante Inhalte gefunden haben.
- **Beliebte Suchbegriffe:** Finden Sie heraus, welche Suchbegriffe bei Ihren Nutzern besonders beliebt sind, um wertvollere Inhalte zu erstellen und UX/UI Ihrer Website zu optimieren.
- **Schlüsselwörter ohne Treffer:** Wenn ein Nutzer bei einer Suchabfrage keine Ergebnisse erhält, bedeutet dies entweder, dass auf Ihrer Website Inhalte fehlen oder dass die Suche nicht genau versteht, wonach die Nutzer suchen.
- **Häufige Rechtschreibfehler:** Welche Suchbegriffe schreiben Ihre Nutzer besonders häufig falsch? Nutzer, die nach „Seute“ anstelle von „Seite“ suchen, erwarten trotzdem eine volle Liste mit Suchergebnissen, unabhängig von ihrem Rechtschreibfehler.

# Monitoring durch Observability

Observability umfasst unter anderem das Monitoring der Anwendungsleistung (Application Performance Monitoring, APM) und ist eine ganzheitliche und dynamische Methode, um herauszufinden, wie Ihre komplexen Systeme funktionieren.

Der Begriff „Observability“ bezieht sich auf die Erfassung und Analyse von Daten, wie etwa Logs und Metriken, um ausführliche Einblicke in das Verhalten der Anwendungen zu liefern, die in Ihren Umgebungen ausgeführt werden. Observability kann auf jedes System angewendet werden, das Sie entwickeln und überwachen möchten.

Für den Umgang mit komplexen Systemen und Unmengen an Logs in listenlosen Daten sollten Sie mit Ihrem DevSecOps-Team zusammenarbeiten, um Probleme proaktiv diagnostizieren, analysieren und zu ihrem Ursprung zurückverfolgen zu können.

## Vorteile von Observability

- Performance-Monitoring für schnelle Lösungen
- Vollständige und sofortige Transparenz
- Eliminierung von Tool-Silos
- Bessere Nutzererlebnisse

# Monitoring der Anwendungsleistung (Application Performance Monitoring, APM)

APM ist ein weiteres Tool, das die fortlaufende Wartung stark erleichtern kann. APM-Lösungen sammeln, überwachen und analysieren Telemetriedaten aus Websites und Diensten, liefern anwendungsübergreifende End-to-End-Transparenz für Entwicklungsteams, um Abhängigkeiten zwischen Anwendungen und Diensten besser verstehen und Fehler oder Leistungsengpässe beheben zu können.

Mit diesen Einblicken können die Teams Probleme proaktiv angehen, anstatt darauf zu warten, dass sich die Nutzer beschweren – oder noch schlimmer, Ihre Website verlassen und nie wieder zurückkommen. Die Entwickler können auch Alerts für Beeinträchtigungen des Nutzererlebnisses einrichten, um fundierte Entscheidungen über potenzielle Verbesserungen zu treffen.

APM-Lösungen speichern und nutzen auch historische Daten, um Trends zu ermitteln und Ausreißer bei Performance-KPIs (Schlüsselkennzahlen) wie etwa Latenz und Durchsatz sowie bei Geschäfts-KPIs zu erkennen.

**Möchten Sie diese Best Practices mit Ihrem DevSecOps-Team teilen**

Observability-Anleitung jetzt herunterladen

## Vorteile von APM

- Mehr Stabilität und Uptime
- Weniger Incidents
- Schnellere Problembehebung
- Hochwertige Softwareveröffentlichungen
- Identifizieren von Infrastrukturverbesserungen
- Mehr Produktivität
- Bessere Nutzererlebnisse
- Umsatzsteigerung

# Community

Freundliche Nachbarn sind ein entscheidender Faktor für unsere Nachbarschaft. Ebenso ist eine florierende und hilfreiche Community aus Software-Nutzern hilfreich, um knifflige Probleme zu lösen, neue Ideen zu entwickeln und den Erfolg Ihrer Sucherlebnisse zu garantieren.

Mit Ihrer Teilnahme an der Elastic-Community erhalten Sie Zugang zu einem Support-Netzwerk, während Sie Ihr Erlebnis mit Elasticsearch erstellen.

Mit einem Platz in der ersten Reihe der Elastic-Community können Sie und Ihr Team auf das gesammelte Wissen anderer Entwickler zugreifen, die sich auf derselben Journey befinden:

- Elastic-Community
- Elastic Contributor Program und Awards
- Elastic-Community-Foren
- Slack-Arbeitsbereich der Elastic-Community
- Elastic-YouTube-Kanal

Wo immer Sie hinschauen, werden Sie sehen, wie Elasticsearch-Profis und Anfänger einander helfen.

Unsere Community basiert seit den Anfangstagen auf drei Produkten (die inzwischen zu einem Erlebnis vereint wurden): Elasticsearch, Logstash und Kibana, oder auch ELK. Der ELK Stack wurde von Ihrem freundlichen Elch Elky vertreten.

# Suche als Bauvorhaben

Wie Sie in dieser Anleitung erfahren haben, ist die Erstellung von Sucherlebnissen für Unternehmen – genau wie der Bau Ihres Traumhauses – ein anspruchsvoller und doch lohnender Prozess.

Elastic hilft Ihnen dabei. Mit einer der branchenweit beliebtesten Such- und Analysemaschinen – Elasticsearch – schlagen wir eine Brücke zwischen LLMs und der Suche, damit Sie benutzerdefinierte generative KI-Anwendungen mit Ihren Unternehmensdaten erstellen können.

Die Elasticsearch Relevance Engine™ (ESRE) unterstützt KI-Lösungen für private Datensätze mit einer Vektordatenbank und Machine-Learning-Modellen für semantische Suchfunktionen, um Entwicklern von Suchanwendungen mehr Relevanz zu liefern. ESRE kombiniert die Stärken von KI mit der textbasierten Suche von Elastic und bietet Entwicklern eine vollständige Suite aus komplexen Abrufalgorithmen und Integrationsmöglichkeiten mit großen Sprachmodellen (Large Language Model, LLM), inklusive des Elastic Learned Sparse Encoder (ELSER). Dieses von Elastic trainierte Modell erfordert keine Feinjustierung mit Ihren eigenen Daten und ist für zahlreiche Anwendungsfälle sofort einsetzbar.

**Sind Sie bereit, eine bessere Suche zu erstellen?**

Quickstart-Anleitung ansehen

# Vielen Dank!