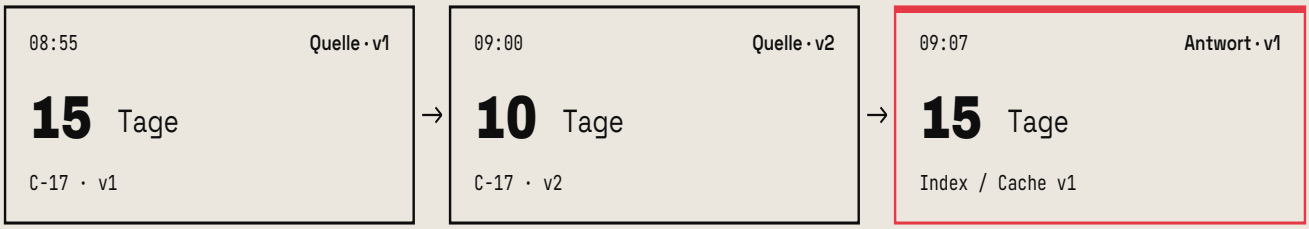
TECHNIKDOSSIER · RAG

Datenbank aktuell, KI-Antwort veraltet: die Aktualität im RAG steuern

Eine aktuelle Quelle garantiert noch keine aktuelle Antwort. Dieser Artikel begleitet das fiktive Dokument C-17 von der Quelle bis zur Antwort und schlägt vier getrennte Übungen vor: Korrektur, Löschung, Berechtigungsentzug und Wiederanlauf.

C-17: die Version verfolgen



Fiktives Beispiel

Vorbereitet am 2. November 2026

Vorgesehene kanonische URL

https://datanalysis.re/ressources/artificial-intelligence/rag-aktualitaet-quelle-index-cache

Das Wichtigste

01 Aktualität ist entlang der gesamten Kette zu prüfen: Quelle, Extraktion, Index, Kontext und Cache.

02 MongoDB Search und Vector Search indexieren asynchron. Die Caches der Anwendung folgen eigenen Regeln.

03 Ein nützlicher Quellenbeleg nennt die tatsächlich verwendete Version, die Passage und den Abrufzeitpunkt.

04 Korrektur, Löschung und Berechtigungsentzug sind unterschiedliche Szenarien. Eine aktuelle Antwort kann für eine Person dennoch unzulässig sein.

05 Beim Wiederanlauf sollen Versuche sichtbar bleiben und ihre Wirkungen kontrolliert werden. Zwei Versuche für eine Wirkung sind hier ein fiktiver Sollwert, keine Produktgarantie.

DER VOLLSTÄNDIGE ARTIKEL

Eine Antwort kann das richtige Dokument zitieren und trotzdem eine veraltete Version verwenden. In einem System mit Retrieval-Augmented Generation, kurz RAG, werden Inhalte abgerufen und anschließend dem Modell bereitgestellt. Die Aktualität hängt von der gesamten Verarbeitungskette ab. Das Änderungsdatum der Quelle liefert einen ersten Anhaltspunkt. Zusätzlich müssen die indexierte Version, der übergebene Kontext und eine gegebenenfalls zwischengespeicherte Antwort nachvollziehbar sein.

Dieser Artikel schlägt ein Verfahren zur Prüfung dieser Kette vor. Das Beispiel ist fiktiv: Es beschreibt erwartete Kontrollen, keinen Kundenvorfall und keinen Test in einer Produktionsumgebung.

Mit einer überprüfbaren Zeitfolge beginnen

Um 08:55 Uhr nennt Version v1 des Dokuments C-17 eine Bearbeitungsfrist von fünfzehn Tagen. Um 09:00 Uhr wird die korrigierte Version v2 veröffentlicht: Die geltende Frist beträgt zehn Tage. Um 09:07 Uhr nennt eine Antwort weiterhin fünfzehn Tage und verweist auf C-17.

In unserem Szenario beruhen der verwendete Index und Cache noch auf v1. Der Dokumentverweis ist somit erkennbar, die Antwort entspricht aber nicht mehr der aktuell geltenden Version.

Bevor Sie den Prompt ändern, sollten Sie den tatsächlich durchlaufenen Verarbeitungspfad feststellen. Welche Extraktion hat den Index gespeist? Welche Version wurde abgerufen? Wurde das Modell aufgerufen oder eine frühere Antwort wiederverwendet? Ohne diese Spuren lassen sich mehrere mögliche Ursachen nicht unterscheiden.

Die an der Antwort beteiligten Kopien erfassen

Wir schlagen eine einfache Bestandsaufnahme vor: fachliche Quelle, extrahierter Text, Textabschnitte, Embeddings, Suchindex, zusammengestellter Kontext, Antwort-Cache und der den Nutzenden angezeigte Verlauf. Jedes Element braucht eine Identität, eine bekannte Version und eine verantwortliche Stelle für seine Aktualisierung.

Diese Übersicht verhindert, dass „die Datenbank“ als ein einziges Objekt behandelt wird. Eine korrigierte Quelle kann neben einem alten Textabschnitt bestehen. Ein aktualisierter Index kann mit einem Cache kombiniert werden, der nach seiner eigenen Ablaufregel noch gültig ist. Ein Verlauf darf eine frühere Antwort als Archiv anzeigen, ohne sie erneut zu einer gültigen Antwort auf eine aktuelle Frage zu machen.

Für jede Kopie ist eine konkrete Frage entscheidend: Welches Ereignis ersetzt sie, erklärt sie für ungültig oder verhindert, dass sie gelesen wird?

Die erforderliche Aktualität am Anwendungsfall ausrichten

MongoDB dokumentiert eine asynchrone Indexierung und Eventual Consistency für MongoDB Search und Vector Search. Eine Änderung in der Collection steht einer Suchabfrage daher nicht zwangsläufig sofort zur Verfügung. 1 2

Diese Eigenschaft muss in die Gestaltung des Dienstes einfließen. Eine interne Dokumentationssammlung und eine dringende betriebliche Anweisung können unterschiedliche Verzögerungen tolerieren. Für jeden Anwendungsfall sind die zulässige Verzögerung und das erwartete Verhalten bei ihrer Überschreitung festzulegen.

Wir empfehlen mehrere Zeitstempel: Änderung der Quelle, Übernahme der Änderung, Aktualisierung der durchsuchbaren Darstellung und Erzeugung der Antwort. Eine einzige Messgröße verdeckt, welcher Schritt hinterherhinkt. Wenn sich die notwendige Aktualität nicht feststellen lässt, kann die Anwendung eine Schlussfolgerung verweigern, eine Prüfung verlangen oder die Einschränkung ausdrücklich anzeigen.

Prüfen Sie auch, wie Embeddings verwaltet werden. Wenn die Anwendung sie erzeugt, ist ihre Erneuerung Teil ihrer Pipeline. Eine Textänderung allein garantiert nicht, dass die Embeddings neu berechnet wurden.

Bedingungen für die Wiederverwendung aus dem Cache festlegen

Ein Cache vermeidet wiederholte Arbeit. Ob eine Antwort wiederverwendet werden darf, sollte jedoch vom Kontext abhängen. Bei C-17 wäre die Frage als alleiniger Cache-Schlüssel unzureichend: Nach Veröffentlichung von v2 kann dieselbe Frage eine andere Antwort erfordern.

Je nach Architektur kann diese Entscheidung Dokumentversionen, den erlaubten Zugriffsumfang, die Version fachlicher Regeln und ein maximales Alter berücksichtigen. Das sind zu prüfende Entwurfsentscheidungen, keine automatischen Eigenschaften der Suchmaschine.

Eine Ablaufzeit begrenzt die Speicherdauer. Sie ersetzt nicht unbedingt die Invalidierung nach einer kritischen Korrektur oder einer Änderung der Berechtigungen. Berücksichtigen Sie außerdem ein verspätetes Invalidierungssignal: Welche Prüfung verhindert dann, dass eine nicht mehr zulässige Antwort ausgegeben wird?

Eine Löschung entlang der gesamten Kette prüfen

Betrachten wir eine eigenständige Variante: Um 09:10 Uhr wird C-17 aus der Quelle gelöscht. Die Prüfung endet nicht mit dem Erfolg dieses Vorgangs. Untersuchen Sie die Kopien, mit denen das Dokument gefunden, der Kontext zusammengestellt und eine Antwort wiederverwendet wird.

MongoDB Change Streams liefern Löschereignisse mit der Identität des Dokuments. Das vollständige Dokument ist im Standardereignis nicht mehr enthalten; unter bestimmten Voraussetzungen lassen sich Pre-Images konfigurieren. 3

Die Anwendung muss deshalb die Abhängigkeiten kennen, über die ihre Textabschnitte und Caches gefunden werden. Wir empfehlen, die Löschung über den vollständigen Nutzungspfad zu prüfen: Suche, umformulierte Frage, Antwort aus dem Cache und Öffnen eines früheren Belegs. Für eine zu Archivzwecken aufbewahrte Kopie müssen ausdrückliche Regeln gelten. Sie darf nicht mit einer Quelle verwechselt werden, die weiterhin für Antworten genutzt werden kann.

Berechtigungen unabhängig von der Aktualität prüfen

Der Entzug einer Berechtigung ist eine weitere, separat zu prüfende Variante. C-17 existiert weiterhin, seine Version v2 ist korrekt, aber die fragende Person darf das Dokument nicht mehr einsehen.

Die Antwort kann damit vollständig aktuell und für diese Person trotzdem unzulässig sein. Zugriffskontrollen müssen sowohl die Inhalte betreffen, die in den Kontext gelangen könnten, als auch bereits gespeicherte Antworten. Die Aufforderung an das Modell, Berechtigungen zu beachten, ersetzt keine von der Anwendung durchgesetzte Autorisierungsentscheidung.

MongoDB beschreibt unter anderem ein Mandantenfeld als Vorfilter für die Suche. 4 Das ist ein nützlicher Mechanismus, wenn die Architektur ihn vorsieht. Das System muss dennoch den berechtigten Mandanten bestimmen und feinere Zugriffsregeln behandeln. Ein von der anfragenden Person frei gewählter Filter bildet für sich genommen keine Zugriffspolitik.

Den Quellenbeleg mit einer Version verbinden

Ein funktionierender Link zeigt, dass sich eine Ressource in einem bestimmten Kontext öffnen lässt. Er beweist nicht, dass die zitierte Passage zur aktuell geltenden Version gehört oder dass die empfangende Person sie lesen darf.

Für einen überprüfbaren Beleg empfehlen wir, die Dokumentkennung, ihre Version, die verwendete Passage und den Abrufzeitpunkt aufzubewahren. Das Provenienzmodell W3C PROV-O unterscheidet unter anderem Zitat, Revision und Invalidierung. 5

Das betriebliche Ziel ist eine präzise Antwort: Auf welchem Zustand des Dokuments beruht diese Schlussfolgerung? Eine neue Version kann eine erneute Prüfung erfordern. Ein früherer Beleg kann weiterhin helfen, eine damalige Entscheidung zu verstehen, sofern sein historischer Kontext klar erkennbar bleibt.

Den Wiederanlauf vorbereiten, ohne Wirkungen zu vervielfachen

Die Aktualität hängt auch vom Neustart nach einer Unterbrechung ab. Ein verarbeitender Dienst kann eine Änderung bereits angewendet haben und stoppen, bevor er seinen Fortschritt speichert. Beim Wiederanlauf begegnet ihm möglicherweise dasselbe Ereignis erneut.

Unsere Übung verwendet das fiktive Ereignis e-417, das C-17 auf v2 umstellen soll. Das Prüforakel der Simulation ist eindeutig: Zwei Versuche sollen eine einzige fachliche Wirkung erzeugen, während beide Versuche in der Ablaufspur sichtbar bleiben. Dies ist das erwartete Ergebnis der Übung, keine MongoDB zugeschriebene Garantie.

Die Change-Streams-Spezifikation ordnet die Verantwortung für idempotente Verarbeitung der Anwendung zu. 6 Definieren Sie eine Ereignisidentität, eine Strategie zur Duplikaterkennung oder einen idempotenten Schreibvorgang. Prüfen Sie anschließend die Grenze zwischen Aktualisierung und Speicherung des Fortschritts. Externe Wirkungen benötigen eine eigene Behandlung.

Ein Resume-Token ersetzt diese Kontrollen nicht. Der Wiederanlauf hängt außerdem von der Historie ab, die noch im Oplog verfügbar ist. 7

Messen, was das Team erklären können muss

Eine Betriebsprüfung kann Verzögerungen je Verarbeitungsschritt, unbearbeitete Änderungen, fehlgeschlagene Invalidierungen und Wiederanläufe mit manuellem Eingriffsbedarf erfassen. Die Schwellenwerte sollten gemeinsam mit den fachlich Verantwortlichen festgelegt und nicht aus einem Beispiel übernommen werden.

Für jeden Alarm ist zu bestimmen, wer eingreift, welche Antworten noch ausgegeben werden dürfen und wie ein überprüfter Zustand wiederhergestellt wird. Bewahren Sie außerdem genügend Spuren auf, um den Weg einer Antwort zu rekonstruieren, ohne Dokumentinhalte unnötig offenzulegen.

Die abschließende Abnahmefrage ist einfach: Können wir zeigen, warum diese Person zu diesem Zeitpunkt genau diese Version erhalten hat? Eine klare Antwort verbindet die technische Umsetzung mit dem fachlichen Bedarf.

Die Verarbeitungskette erproben

Beginnen Sie mit einem fiktiven Dokument und einer unveränderten Frage. Variieren Sie jeweils nur eine Bedingung: Korrektur, Löschung, Berechtigungsentzug und anschließend Wiederanlauf der Verarbeitung. Halten Sie für jedes Szenario das erwartete Ergebnis vor der Ausführung fest und dokumentieren Sie das beobachtete Ergebnis separat.

So entsteht eine reproduzierbare Prüfung von der Quelle bis zur Antwort, einschließlich der zu behebenden Abweichungen und der anstehenden Entscheidungen.

Bringen Sie für eine solche Prüfung mit DATANALYSIS ein Schema Ihrer Pipeline und ein Beispieldokument mit, dessen Aktualität für eine Entscheidung wichtig ist: eine Prüfung der RAG-Aktualität vorbereiten.

Fragen & Antworten

Warum kann eine RAG-Antwort trotz aktueller Quelle veraltet sein?

Extraktion, Index oder Cache können noch eine ältere Version verwenden. MongoDB Search und Vector Search arbeiten mit Eventual Consistency. Verfolgen Sie die Version je Schritt und prüfen Sie, ob die Antwort neu erzeugt oder wiederverwendet wurde.

Garantiert eine Ablaufzeit des Caches aktuelle Antworten?

Die Ablaufzeit begrenzt die Aufbewahrung. Eine kritische Korrektur oder geänderte Berechtigungen können jedoch eine frühere Invalidierung erfordern. Das nötige Verhalten hängt vom Anwendungsfall ab und muss von der Anwendung definiert und geprüft werden.

Sind Löschung und Berechtigungsentzug derselbe Testfall?

Nein. Eine Löschung entfernt das Quelldokument. Ein Berechtigungsentzug nimmt einer Person das Zugriffsrecht, obwohl das Dokument weiter existieren kann. Kopien, Belege und gespeicherte Antworten sind für beide Szenarien getrennt zu prüfen.

Welche Angaben gehören zu einem Quellenbeleg?

Dokumentkennung, Version, verwendete Passage und Abrufzeitpunkt. Ein funktionierender Link belegt weder die Aktualität der Passage noch die Leseberechtigung der empfangenden Person.

Was beweist die Übung e-417 mit zwei Versuchen?

Sie definiert einen Sollwert: zwei sichtbare Versuche und eine fachliche Wirkung. Sie beweist nicht, dass ein reales System diesen Wert bereits erreicht, und schreibt MongoDB keine Garantie für genau einmalige Ausführung zu.

Mit welchem Test sollte die Aktualitätsprüfung beginnen?

Wählen Sie ein fiktives Dokument und eine unveränderte Frage. Halten Sie das erwartete Ergebnis fest, ändern Sie eine Bedingung und dokumentieren Sie das beobachtete Ergebnis. Beginnen Sie mit einer Korrektur; prüfen Sie Löschung, Berechtigungsentzug und Wiederanlauf anschließend separat.

Quellen

Dokumentation geprüft am 2. November 2026.

1. MongoDB Search: Indexierung und Eventual Consistency
2. MongoDB Vector Search: Hinweise zu Indizes
3. MongoDB: Löschereignisse und Pre-Images
4. MongoDB: Mandantenarchitektur und Vorfilter
5. W3C: PROV-O, die Provenienzontologie
6. MongoDB: Change-Streams-Spezifikation und Idempotenz der Anwendung
7. MongoDB: Wiederaufnahme von Change Streams