Внедрение AWS GraphRAG сокращает циклы разработки лекарств на 87%

Источник: AI News Healthcare

Оригинал: https://www.artificialintelligence-news.com/news/aws-graphrag-deployment-cuts-drug-research-cycles-by-87/

GraphRAG | NLP | облачные технологии | обработка данных | разработка лекарств | фармацевтика

Недавнее развертывание **GraphRAG** на базе **AWS** сократило циклы исследований и разработок лекарственных препаратов в фармацевтической среде на 87 процентов. Это ускорение достигается путем интеграции ранее разделенных проприетарных баз данных в единый и доступный для запросов **граф знаний** (knowledge graph).

Исторически сложилось так, что начальные фазы сбора и скрининга данных занимали более шести месяцев на одну итерацию, обеспечивая низкий показатель успеха — всего пять процентов. Критически важные наборы данных — от специфических для данной области клинических показателей до внутренних инженерных и лабораторных записей — были изолированы в различных средах хранения, что фактически мешало специалистам по данным (data scientists) выявлять скрытые корреляции. При увольнении сотрудников они уносили с собой важнейший контекст проектов, что приводило к остановке активных исследований.

Компания **AWS** разработала решение для соединения этих систем, объединив графовые базы данных с технологиями **NLP** (Natural Language Processing — обработка естественного языка).

Конфигурация опирается на фреймворк **GraphRAG** и использует **Amazon Neptune Analytics** и **Bedrock**, чтобы превратить разрозненные точки данных в поисковую сеть. Пользователи могут вводить стандартные запросы на естественном языке и получать ответы, сопоставленные с проверенной профильной литературой и внутренними наборами данных.

Однако объединение изолированных проприетарных наборов данных с неструктурированными репозиториями открытого доступа по-прежнему создает значительные проблемы с нормализацией данных, что требует строгого управления схемами для предотвращения неточного реляционного картирования и снижения риска «галлюцинаций» ИИ.

Построение графа знаний

Компании могут подключать собственные графы знаний. Система подтягивает разрозненные неструктурированные файлы из публичных баз данных, таких как **PubMed**, и смешивает их с внутренними корпоративными записями. Инструменты, такие как **Amazon Comprehend Medical**, сканируют этот текст для извлечения стандартных медицинских кодов. **Amazon Bedrock**, работающий на базе модели **Anthropic Claude 4.5 Sonnet**, резюмирует содержимое документов и определяет тематическую релевантность.

Затем функции **AWS Lambda** и пакетная загрузка через **Amazon S3** направляют эти обработанные элементы в **Amazon Neptune Analytics**. Полученный граф знаний структурирует данные в отдельные узлы, представляющие основные сущности, такие как специфические для области классы, авторы, исходные журналы и фрагменты встроенного текста. Ребра графа определяют отношения между этими узлами, выстраивая иерархические классификации и ассоциации сущностей. Это структурированное представление обеспечивает детерминированную основу, необходимую для точного извлечения информации.

Схема базы данных устанавливает строгие границы процесса обнаружения через **RAG** (Retrieval-Augmented Generation — генерация с дополнением извлечением). Узлы структурированы таким образом, чтобы фиксировать конкретные состояния и иерархически сопоставлять их с установленными онтологиями, в то время как узлы авторов и журналов обеспечивают прослеживаемость опубликованных исследований. Длинные документы разбиваются на удобочитаемые текстовые сегменты с использованием

стратегий сегментации (chunking) **Amazon Bedrock Knowledge Base**, а специфические узлы классификации привязывают неструктурированные текстовые данные к стандартизированным диагностическим показателям.

Эксплуатация этой графовой архитектуры требует выделения определенных облачных ресурсов. Стандартный граф **Amazon Neptune Analytics**, работающий с 16 выделенными единицами памяти, несет операционные расходы в размере 0,48 доллара США в час. Среды разработки, такие как блокноты **Amazon SageMaker Jupyter**, работающие на инстансах **t3.medium**, добавляют базовые расходы на вычисления и хранение. Организации также должны учитывать динамические затраты на потребление токенов, генерируемые моделью **Amazon Bedrock Claude 4.5 Sonnet** во время обработки запросов и генерации аннотаций.

Инструментарий **GraphRAG** выступает в качестве исполнительного уровня между пользовательским интерфейсом и базовой базой данных. Специализированный компонент **Knowledge Graph Linker** обрабатывает входящие запросы на естественном языке, извлекает соответствующие сущности, используя нечеткое строковое индексирование, и сопоставляет их с установленными узлами графа. Система проходит по путям сети, чтобы создать правдоподобные реляционные связи, прежде чем сформулировать ответ через языковую модель, размещенную в **Bedrock**.

Точность извлечения зависит от конфигурации сопоставления сущностей. Компонент **EntityLinker** сопоставляет термины естественного языка из запросов пользователей со структурированной схемой данных. Этот процесс нечеткого сопоставления (fuzzy matching) справляется с присущим шумом и разнообразной терминологией, встречающейся в сложных корпоративных наборах данных, гарантируя, что пользователи извлекают правильные узлы, даже если используют неточные формулировки.

Модульность и архитектура системы

Извлечение данных в значительной степени опирается на специализированный ИИ-парсинг; архитектура использует **Claude** для оценки исходных документов и генерации кратких аннотаций. Затем специализированные инструменты сопоставляют эти сложные текстовые описания со стандартизированными таксономиями.

Инструментарий **GraphRAG Python** инициализирует **BedrockGenerator** для обеспечения взаимодействия на естественном языке, в то время как инженеры настраивают компонент **Knowledge Graph Linker**, чтобы связать

хранилище графа с языковой моделью. Эта интеграция создает прямой интерфейс для выполнения запросов и генерации ответов, строго основанных на доступных данных графа.

Архитектура разделяет три основные функции: инициализацию языковой модели, интерфейс графа и связывание сущностей. Благодаря модульности системы команды могут заменять языковую модель или корректировать структуру графа без необходимости демонтажа и пересборки всего приложения.

Активные развертывания архитектуры **Neptune** и **Bedrock** возвращают точные, проверяемые цитаты для каждого сгенерированного ответа. Система отображает весь путь рассуждения, демонстрируя конкретные шаги обхода графа, использованные для достижения вывода.

Ключевые показатели эффективности первых корпоративных пользователей включают сокращение длительности исследовательского цикла на 87 процентов. Начальные фазы обнаружения, которые ранее требовали шесть месяцев, теперь завершаются за три недели, а скорость извлечения данных улучшилась на 85 процентов, что напрямую способствует более быстрому тестированию гипотез. Кроме того, время проверки исследований сокращается на 70 процентов благодаря функциям автоматического сопоставления цитат и верификации источников.

Инженерные группы могут интегрировать новые публичные базы данных или внутренние заметки в существующую структуру графа без нарушения работы активных интерфейсов запросов. Для обеспечения управления и соблюдения нормативных требований фиксируются точные доказательные цепочки, необходимые для подачи документов в регулирующие органы, а визуализация обхода графа наглядно показывает, как модель ИИ связала сложные переменные. Команды могут проследить каждый результат непосредственно до исходных документов, выполняя требования комплаенса в отношении научной целостности.

Наконец, поддержание централизованного графа знаний предотвращает деградацию данных. Когда ведущие ученые увольняются, их неявные знания о поведении системы или неудачных экспериментах остаются индексированными в базе данных **Neptune**. Новый персонал может запрашивать систему для изучения прошлых решений и мгновенно получать доступ к историческому контексту текущего проекта.

По мере созревания фреймворков **GraphRAG** маловероятно, что эта модель развертывания останется ограниченной только фармацевтическими исследованиями. Способность детерминированно сопоставлять внутренние неструктурированные данные с проверенными публичными репозиториями представляет собой шаблон для любого предприятия, стремящегося извлечь полезную информацию из фрагментированных устаревших систем.

См. также: **Insilico Medicine** продвигает препарат на базе ИИ для лечения ИЛФ (IPF) к III фазе клинических испытаний.
