10 ноября 2025

LLM и дата-каталог: описание метаданных

Как описать метаданные и поддерживать дата-каталог в актуальном состоянии? И как при этом может помочь LLM?

Согласно исследованию, 93% организаций, если еще не внедрили дата-каталог, как инструмент, поддерживающий процессы Data Governance, то по крайней мере об этом задумываются. Процессное управление метаданными, как бизнес, так и техническими, значительно снижает T2M (time-to-market) для дата-продуктов, повышает операционную эффективность за счет уменьшения дублирования данных и использования единой методологии расчета показателей. Но эти эффекты достижимы в полной мере только при наличии описаний или бизнес-контекста у метаданных. Так, один из важных вопросов при внедрении дата-каталога: как выполнить инициализирующее описание метаданных и поддерживать его в актуальном состоянии?

Как подойти к вопросу?

Разумеется, можно делать это вручную, силами дата-стюардов, аналитиков или привлекать другие роли, но для этого необходима процессная модель и затраты на привлечение команды. Ниже опишем подходы, которые позволяют добиться одновременно и высокой доли описания метаданных, и минимизации расходов на работу с ними.

Описание метаданных в дата-каталоге — это типовая задача генерации контента для LLM (large language model). Модели необходимо дать контекст: в рамках какого домена и/или предметной области реализованы объекты технических метаданных, что понимается под доменом и предметной областью в организации, каковы их границы, какая принята терминология и прочее. Обобщая – дать семантику, в границах которой необходимо сгенерировать описания для метаданных.

Для этого обратимся к функциональной архитектуре дата-каталога: подавляющее большинство из них содержит два основных модуля – каталог технических метаданных и бизнес-глоссарий. Именно последний (при условии его ведения), является превосходным источником бизнес-контекста организации.

Зачем нужен RAG?

Для учета бизнес-контекста в задачах с применением LLM существует архитектурный подход RAG (Retrieval Augmented Generation) – генерация с дополнением выборки из внешних источников данных, которая и дает LLM тот самый контекст или семантику. Упрощенно, в RAG подходе языковая модель учитывает не только непосредственно переданный промт, но и дополнительный контекст, полученный из внешнего источника. Обычно в качестве внешнего источника предполагается векторная база данных, основное назначение которой в RAG-подходе — это индексация содержания и обеспечение выдачи результата по запросу, соответствующему контексту. Но большинство дата-каталогов, доступных на отечественном рынке, в качестве компонента содержат OpenSearch/Elasticsearch, которые, по сути, обеспечивают те же самые функции в рамках дата-каталога.

При этом компоненты OpenSearch/Elasticsearch для поиска в дата-каталогах индексируют в том числе и данные бизнес-глоссария, а использование k-NN плагина позволяет реализовать не только лексический, но и семантический поиск по содержимому.

Таким образом, для автоматизации описания технических метаданных нам необходима, собственно, сама LLM и контекст бизнес-глоссария, который языковая модель может получать через интеграцию с компонентами дата-каталога OpenSearch/Elasticsearch. Такой подход позволит получить автоматизированное описание метаданных в дата-каталоге и в то же время это менее трудозатратный способ реализации за счет использования уже готового функционала индексации.

Где взять данные?

Но возникает вопрос, как LLM может получить тот самый контекст, если по каким-то причинам в дата-каталоге не ведется бизнес-глоссарий или он ведется ситуативно и не отличается полнотой наполнения. Разумеется, в первую очередь рекомендуем вести бизнес-глоссарий, так как именно он позволяет обеспечить единую методологию расчета показателей и вовлечь бизнес-пользователей в процессы работы с данными. Если по каким-то причинам использовать его невозможно или недостаточно для полного бизнес-контекста, то нужную семантику могут дать различные технические задания, спецификации, S2T (source to target) маппинги, интеграционные контракты и прочие документы, описывающие реализацию той или иной базы данных или другого источника метаданных.

В этом случае реализацию индексации и поиска потребуется обеспечить самостоятельно, что будет сложнее, так как потребуется извлечь текст из документов различных форматов, определить подход к «чанкованию» (синтаксическое разбиение текста на фрагменты) и реализовать поисковые индексы для полученных фрагментов. Конечно, возможен и гибридный подход, при котором источником для контекста выступают и внешние по отношению к дата-каталогу документы, и бизнес-глоссарий.

Вне зависимости от того, какой подход для генерации описаний метаданных будет применен, их наличие в дата-каталоге позволяет сократить T2M за счет снижения затрат на поиск, способствует переиспользованию данных в организации, стимулирует применение self-service анализа и упрощает взаимодействие между участниками процессов использования данных.