УДК 004.75

МЕТОДЫ ХРАНЕНИЯ, АГРЕГАЦИИ И АНАЛИЗА ГЕТЕРОГЕННЫХ ДАННЫХ НА ПРИМЕРЕ АРТЕФАКТОВ РАЗРАБОТКИ

Курочкин Д. М.[1], Нестеренко Н. В.[1], Ханалайнен Д. М.[1]
Научный руководитель – канд. техн. наук, доцент Перл И. А.[1]
[1]Университет ИТМО

Введение

В современных процессах разработки программного обеспечения данные о ходе работ и артефактах распределены по множеству инструментов (репозитории, трекеры задач, wiki и др.), что усложняет их последующую обработку и анализ. Это приводит к высокой стоимости ручной навигации по данным, проблемам сопоставления объектов из разных источников и невозможности быстро получать “целостную картину” проекта. Целью данного исследования является решение проблемы обработки и хранения данных из разрозненных источников путем реализации нового программного комплекса.

Основная часть

В рамках исследования была спроектирована и прототипирована архитектура системы агрегации данных и базовый программный интерфейс управления организационной структурой (домены/проекты/источники данных), включая REST API и проработку подходов к масштабированию и тестированию. Для обеспечения безопасного доступа к ресурсам системы разработана и реализована иерархическая ролевая модель на основе Keycloak. Модель учитывает иерархию ресурсов Domain, Project, Pump (источник данных) и вводит роли Administrator, DomainOwner, ProjectOwner, Developer; роли отражают уровень ответственности и используются серверным компонентом. Для унифицированного хранения разнородных данных спроектирована архитектура хранилища в Tarantool в виде канонической модели Entity/Attribute/Update/Relation/Domain. Связи между сущностями задаются явно, что закладывает основу для различных вариантов дальнейшего представления данных и аналитики. Также проработан событийный контур обработки: данные публикуются в RabbitMQ и потребляются модулем connector на стороне Tarantool, что позволяет поддерживать разные форматы входных данных при сохранении единого контракта хранения. Наконец, спроектирована аналитическая подсистема, ориентированная на семантический поиск и выявление похожих сущностей (kNN top-K) на основе векторных представлений (embeddings) и векторного индекса; отдельно выделены сценарии поиска по сущности, поиска по текстовому запросу и формирования кандидатов на дедупликацию. В качестве векторной БД планируется использовать Qdrant.

Выводы

В результате исследования сформирован фундамент для реализации проекта, а именно: разработанной архитектуры и диаграмм для ряда компонентов (в частности, серверной составляющей, СУБД, аналитической подсистемы), реализованы некоторые части системы на языках программирования C#, Python. Разработанное решение позволяет задействовать новые методы в контроле процессов разработки в распределенных командах, а также внедрить программный комплекс в существующие методологии.

Литература

1. Jodavi M., Tsantalis N. Accurate method and variable tracking in commit history //Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. – 2022. – С. 183-195. (Дата обращения 15.02.2026).
2. Метод хранения векторных представлений в сжатом виде с применением кластеризации [Электронный ресурс] // – Режим доступа: https://cyberleninka.ru/article/n/metod-hraneniya-vektornyh- predstavleniy-v-szhatom-vide-s-primeneniem-klasterizatsii/ - Дата обращения: 15.02.2026.
3. Yokomori R., Inoue K. An Empirical Analysis of Git Commit Logs for Potential Inconsistency in Code Clones //arXiv preprint arXiv:2409.08555. – 2024.
4. Vera H. et al. Data modeling for NoSQL document-oriented databases //CEUR Workshop Proceedings. – 2015. – Т. 1478. – С. 129-135.
5. Accurate method and variable tracking in commit history [Электронныйресурс]. — URL: https://www.researchgate.net/publication/365268730_Accurate_method_and_variable_tracking_in_commit_history (дата обращения 15.02.2026)
6. In-Memory Database Systems - A Paradigm Shift [Электронный ресурс]. — URL: https://www.researchgate.net/publication/260089351_In-Memory_Database_Systems_-_A_Paradigm_Shift (дата обращения 15.02.2026)