УДК 004.652:641.1

РЕАЛИЗАЦИЯ ГИБРИДНОЙ МОДЕЛИ ДАННЫХ НУТРИЕНТНОГО ПРОФИЛЯ ПРОДОВОЛЬСТВЕННОГО СЫРЬЯ

Лемешевский Андрей Васильеивч
Шумак Виктор Викторович, д.с.-х.н., доцент
Полесский государственный университет

IMPLEMENTATION OF A HYBRID DATA MODEL FOR THE NUTRIENT PROFILE OF FOOD RAW MATERIALS

Lemeshevsky Andrey, radicalhydrogen@mail.ru
Shumak Viktor, Doctor of Agricultural Sciences, Associate Professor, shumak.v@polessu.by
Polessky State University

Аннотация. Предложена гибридная модель данных для информационной системы мониторинга нутриентного профиля продовольственного сырья, сочетающая реляционное ядро с документоориентированным хранилищем на основе типа JSONB в СУБД PostgreSQL. Разработаны алгоритм многоуровневой валидации и механизм семантической интеграции с классификатором FoodEx2.

Ключевые слова: гибридная модель данных, нутриентный профиль, polyglot persistence, PostgreSQL, JSONB, FoodEx2, мониторинг продовольственного сырья, ETL-конвейер.

Abstract. A hybrid data model for a food raw material nutrient profile monitoring information system is proposed, combining a relational core with a document-oriented storage based on the PostgreSQL JSONB data type. A multi-level validation algorithm and a semantic integration mechanism with the FoodEx2 classifier have been developed.

Keywords: hybrid data model, nutrient profile, polyglot persistence, PostgreSQL, JSONB, FoodEx2, food raw material monitoring, ETL pipeline.

Информационные системы мониторинга нутриентного профиля продовольственного сырья оперируют данными высокой степени гетерогенности: от строго типизированных лабораторных показателей содержания макронутриентов до слабоструктурированных сведений о способах переработки, условиях отбора проб и результатах инструментального анализа. Библиометрический анализ, проведённый A. Yeung [1], показал, что значительная часть существующих баз данных химического состава пищевых продуктов (Food Composition Databases, FCDB) характеризуется ограниченным объёмом записей и построена на статических реляционных схемах, что затрудняет их расширение при включении новых аналитических параметров. Классическое реляционное моделирование при описании нескольких сотен нутриентов, из которых для конкретного продукта фиксируются, как правило, несколько десятков показателей [1], приводит к формированию разреженных таблиц с преобладанием пустых значений. Альтернативная модель Entity-Attribute-Value (EAV), применяемая в ряде зарубежных FCDB, устраняет проблему разреженности, однако порождает деградацию производительности при аналитических запросах вследствие множественных операций соединения таблиц. Указанные ограничения определили направление настоящего исследования — разработку гибридной модели данных, сочетающей строгость реляционной схемы для ядра системы с гибкостью документоориентированного представления для вариативных компонентов нутриентного профиля.

Концепция polyglot persistence, предполагающая применение нескольких технологий хранения в рамках одной информационной системы, получила формальное обоснование в работе N. Roy-Hubara и соавторов [2], где предложена методология выбора оптимальной комбинации СУБД под требования конкретного приложения. В контексте мониторинга нутриентного профиля данный подход реализован посредством объединения реляционного ядра и документоориентированного слоя в единой СУБД PostgreSQL. Принципиальным решением, составляющим элемент научной новизны настоящей работы, является отказ от развёртывания отдельного NoSQL-хранилища в пользу встроенного типа данных JSONB, что позволяет сохранить транзакционную целостность и снизить эксплуатационную сложность при одновременном обеспечении гибкости схемы для вариативных данных.

Реляционное ядро предложенной модели охватывает сущности с устойчивой структурой и высокими требованиями к типовой безопасности. Центральной является таблица food_item, содержащая идентификационные сведения о продукте: наименование, код по ТН ВЭД, привязку к национальным стандартам ГОСТ и СТБ. Таблица sample хранит метаданные лабораторных проб – дату отбора, наименование лаборатории, применённый метод анализа – и связана с food_item через внешний ключ. Справочник нутриентов nutrient_definition обеспечивает привязку каждого показателя к международным кодам INFOODS tagnames, что является предпосылкой для межсистемного обмена данными. Для базовых макронутриентов (энергетическая ценность, белки, жиры, углеводы, вода, зола – показатели, присутствующие практически в каждом описании продукта) применена «широкая» таблица core_nutrients с типизированными столбцами, что обеспечивает выполнение аналитических запросов без каскадных операций соединения. Отслеживание изменений нутриентного профиля при поступлении обновлённых лабораторных данных реализовано по паттерну Slowly Changing Dimension Type 2 (SCD2) с полями valid_from, valid_to и is_current, позволяющему фиксировать хронологию пересмотра состава продукта и проводить ретроспективные эпидемиологические исследования. Опыт проектирования гибридных аналитических хранилищ, описанный С. Л. Подвальным и соавторами [3], подтверждает целесообразность размещения структурированных нормативно-справочных данных в реляционном сегменте для обеспечения строгой транзакционности.

Для расширенного нутриентного профиля – витаминов, минеральных веществ, аминокислот, жирнокислотного состава, фитохимических соединений – предложено хранение в столбце типа JSONB таблицы extended_profile. Каждый документ представляет собой иерархическую структуру с вложенными объектами по категориям нутриентов, где для каждого показателя фиксируются значение, единица измерения, код аналитического метода и индикатор качества данных. Подобная организация устраняет проблему разреженности: в документе присутствуют только фактически определённые параметры, а добавление новых показателей при внедрении современных аналитических методик не требует модификации DDL-схемы базы данных. Для обеспечения поиска по вложенным атрибутам документов задействованы GIN-индексы (Generalized Inverted Index) с операторным классом jsonb_path_ops, позволяющие выполнять выборку продуктов по содержанию конкретного нутриента без полного сканирования коллекции. Согласно официальной документации PostgreSQL, операторный класс jsonb_path_ops обеспечивает более компактное представление индекса и повышенную специфичность поиска по сравнению со стандартным классом jsonb_ops, что существенно для работы с крупными массивами нутриентных данных. В таблице 1 представлена сравнительная характеристика рассмотренных подходов к хранению.

Таблица 1. – Сравнительная характеристика подходов к хранению нутриентных данных

Критерий	Широкая реляционная таблица	EAV-модель	JSONB (предложенная модель)
Гибкость схемы	Низкая	Высокая	Высокая
Типовая безопасность	Полная	Ограниченная	На уровне приложения
Скорость аналитических запросов	Высокая	Низкая	Высокая
Проблема разреженности	Есть	Нет	Нет
Индексирование вложенных атрибутов	Нативное	Через JOIN	GIN
Транзакционная целостность	Полная	Полная	Полная

Выбор JSONB в рамках единой СУБД PostgreSQL вместо развёртывания отдельного документоориентированного хранилища обусловлен рядом факторов. Прежде всего, сохраняется транзакционная целостность: запись в реляционные таблицы и обновление JSONB-документа выполняются в рамках одной ACID-транзакции, что критически важно при загрузке лабораторных данных,

где целостность связи между метаданными пробы и результатами анализа является обязательным условием достоверности. Кроме того, снижается эксплуатационная сложность: отсутствие необходимости в синхронизации между двумя СУБД устраняет класс ошибок, связанных с межсистемной согласованностью данных. Как отмечено А. И. Балесом [4], в микросервисной архитектуре унификация модели данных на уровне справочников позволяет обеспечить согласованность при децентрализованном управлении, что применимо и к случаю объединения реляционных и документоориентированных данных в едином хранилище.

Загрузка данных из гетерогенных источников – лабораторных информационных систем (LIMS), международных баз EuroFIR и USDA FoodData Central, файлов производителей – реализована посредством ETL-конвейера, функционирующего в три стадии. На этапе извлечения (Extract) адаптеры для каждого типа источника обеспечивают приём данных через REST API или из табличных файлов с промежуточным сохранением в staging-области в формате JSON. На этапе трансформации (Transform) выполняется нормализация единиц измерения к стандартной базе «на 100 г съедобной части», маппинг наименований нутриентов на международные коды INFOODS tagnames через справочник синонимов, а также проверка баланса масс: сумма содержания белков, жиров, углеводов, воды и золы должна находиться в диапазоне 97–103 г на 100 г продукта в соответствии с процедурами контроля качества, применяемыми в международных проектах гармонизации FCDB [5]. Опыт европейского проекта Stance4Health [5], в ходе которого была создана унифицированная база данных состава пищевых продуктов путём гармонизации десяти национальных таблиц химического состава с применением кодирования FoodEx2 и INFOODS tagnames, подтверждает необходимость строгой стандартизации на этапе трансформации для обеспечения сопоставимости данных из различных источников. На этапе загрузки (Load) валидированные записи атомарно фиксируются в реляционном ядре и JSONB-столбце в рамках единой транзакции с ведением журнала, фиксирующего источник, дату загрузки и результаты валидации.

В рамках настоящей работы предложен механизм многоуровневой валидации поступающих данных, включающий три уровня контроля. Синтаксический уровень обеспечивает проверку структуры входного пакета на соответствие заданной JSON Schema – контролируются наличие обязательных полей, корректность типов данных и допустимость значений перечислений. Семантический уровень реализует проверку предметных бизнес-правил: контроль диапазонов допустимых значений для каждого нутриента, а также верификацию баланса масс проксимальных компонентов. Классификационный уровень обеспечивает проверку валидности присвоенного кода FoodEx2 и совместимости фасетных дескрипторов путём обращения к актуальному справочнику классификатора. Пакеты данных, не прошедшие валидацию, направляются в карантинную область для последующей верификации оператором с формированием структурированного отчёта о выявленных несоответствиях (таблица 2).

Таблица 2. – Уровни валидации данных

Уровень	Механизм	Пример	Действие
Синтаксический	JSON Schema	Обязательное поле	Отклонение
Семантический	Бизнес-правила	Баланс масс	Карантин
Классификационный	FoodEx2	Проверка кода	Возврат

Семантическая интеграция с международным классификатором FoodEx2, разработанным Европейским агентством по безопасности пищевых продуктов (EFSA), реализована посредством оригинальной двухуровневой схемы хранения. Базовый термин (base term), определяющий принадлежность продукта к одной из категорий иерархии FoodEx2, хранится как внешний ключ в реляционном ядре, что обеспечивает ссылочную целостность и возможность эффективного агрегирования данных по товарным группам. Фасетные дескрипторы — способ кулинарной обработки, тип упаковки, часть растения и иные уточняющие признаки — представлены в виде массива кодов в JSONB-поле, что позволяет динамически формировать полный составной код FoodEx2 при экспорте данных без ограничения количества применяемых фасетов. Подобное разделение составляет элемент научной новизны работы: в ряде существующих реализаций FCDB [1] коды FoodEx2 хра-

нятся либо целиком в текстовом поле, либо разбиваются по фиксированному числу столбцов, что ограничивает расширяемость. Для новых продуктов, поступающих из лабораторий с произвольными текстовыми наименованиями, разработан алгоритм полуавтоматического предложения кода FoodEx2 на основе токенизации наименования и поиска ближайшего термина в словаре классификатора с применением метрики расстояния Левенштейна, после чего оператор подтверждает или корректирует предложенный код.

По итогам проведённого исследования предложена гибридная модель данных, специализированная для предметной области нутрициологического мониторинга и сочетающая реляционное ядро для структурированных метаданных и макронутриентов с JSONB-хранилищем для вариативного расширенного профиля. Разработаны ETL-конвейер с трёхуровневой валидацией и двухуровневый механизм хранения классификации FoodEx2. В отличие от типовых реализаций polyglot persistence, предполагающих развёртывание нескольких СУБД [2], предложенная модель объединяет обе парадигмы в рамках единой СУБД, что устраняет проблему межсистемной синхронизации.

Практическая значимость работы определяется возможностью применения предложенной модели при создании национальной системы мониторинга нутриентного профиля продовольственного сырья Республики Беларусь. Направлениями дальнейшего развития являются внедрение алгоритмов машинного обучения для прогнозирования нутриентного профиля новых продуктов на основе рецептурного состава, интеграция с системами прослеживаемости продовольственной цепочки, а также расширение модели данных для хранения результатов мониторинга контаминантов.

Социальная значимость заключается в создании эксклюзивных товарных продуктов высокого качества под гастрономические запросы потребителей.

Список использованных источников

1. Yeung, A. W. K. Food Composition Databases (FCDBs): A Bibliometric Analysis / A. W. K. Yeung // Nutrients. – 2023. – Vol. 15, no. 16. – Art. 3548. – URL: https://doi.org/10.3390/nu15163548 (дата обращения: 06.04.2026).

2. Roy-Hubara, N. Selecting databases for Polyglot Persistence applications / N. Roy-Hubara, P. Shoval, A. Sturm // Data & Knowledge Engineering. – 2022. – Vol. 137. – Art. 101950. – URL: https://doi.org/10.1016/j.datak.2021.101950 (дата обращения: 06.04.2026).

3. Подвальный, С. Л. Целевая архитектура гибридного аналитического хранилища данных для предприятия электронной коммерции / С. Л. Подвальный, В. Ф. Барабанов, Ф. Г. Логинов, С. А. Коваленко // Вестник Воронежского государственного технического университета. – 2019. – Т. 15, № 4. – С. 19–29. – URL: https://cyberleninka.ru/article/n/tselevaya-arhitektura-gibridnogo-analiticheskogo-hranilischa-dannyh-dlya-predpriyatiya-elektronnoy-kommertsii (дата обращения: 05.04.2026).

4. Балес, А. И. Унифицированная модель данных и её применение в микросервисной архитектуре / А. И. Балес // Современные информационные технологии и ИТ-образование. – 2020. – Т. 16, № 2. – С. 416–425. – URL: https://cyberleninka.ru/article/n/unifitsirovannaya-model-dannyh-i-eyo-primenenie-v-mikroservisnoy-arhitekture (дата обращения: 05.04.2026).

5. Hinojosa-Nogueira, D. Development of an Unified Food Composition Database for the European Project "Stance4Health" / D. Hinojosa-Nogueira, S. Perez-Burillo, B. Navajas-Porras [et al.] // Nutrients. – 2021. - Vol. 13, no. 12. – Art. 4206. – URL: https://doi.org/10.3390/nu13124206 (дата обращения: 06.04.2026).