

**Власов Егор Александрович**

*студент IV курса Российской академии народного хозяйства и государственной службы при Президенте Российской Федерации (Северо-Западный институт управления РАНХиГС), Санкт-Петербург, Российская Федерация*

**ПРОЕКТИРОВАНИЕ ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ  
МОНИТОРИНГА ПРЕДОСТАВЛЕНИЯ  
ГОСУДАРСТВЕННЫХ УСЛУГ НА ОСНОВЕ ГИБРИДНОЙ  
АРХИТЕКТУРЫ ХРАНИЛИЩА ДАННЫХ**

**Аннотация.** В статье рассматривается проблема обработки и анализа массивов транзакционных логов инфраструктуры электронного правительства. Для преодоления ограничений реляционных СУБД при работе с OLAP-нагрузками предложена архитектура гибридного хранилища данных (DWH), физически разделяющая нормативно-справочную информацию (PostgreSQL) и транзакционные факты (колоночная СУБД ClickHouse). Описан процесс проектирования денормализованной инфологической модели по схеме «Звезда» с транзакционной зернистостью, а также организация идемпотентного ETL-конвейера с применением Apache Airflow. В качестве аналитического ядра системы разработан математический аппарат, включающий метод статистического контроля технической доступности шлюзов на базе динамического Z-индекса и предиктивную модель вычисления риска нарушения регламентных сроков (SLA). Реализация предложенного технологического стека и алгоритмов обеспечивает высокую производительность BI-систем и формирует инструментарий для проактивного управления качеством предоставления государственных услуг.

**Ключевые слова:** бизнес-аналитика, электронное правительство, хранилище данных, гибридная архитектура, ETL-конвейер, СУБД ClickHouse, предиктивный мониторинг, Z-оценка, регламентные сроки

**Annotation.** The article addresses the problem of processing and analyzing large volumes of transaction logs generated within e-government infrastructure. To overcome the limitations of relational database management systems in handling OLAP workloads, a hybrid data warehouse (DWH) architecture is proposed that physically separates reference data stored in PostgreSQL from transactional facts stored in ClickHouse, a column-oriented DBMS. The paper describes the design of a denormalized dimensional model based on a star schema at the transactional grain, as well as the orchestration of an idempotent ETL pipeline using Apache Airflow. The analytical core of the system comprises a mathematical framework that includes a statistical control method for monitoring the technical availability of gateways based on a dynamic Z-score and a predictive model for estimating the risk of non-compliance with Service Level Agreement (SLA) requirements. The implementation of the proposed technology stack and analytical methods enables high-performance BI analytics and provides a toolset for proactive management of public service quality.

**Keywords:** business intelligence; e-government; data warehouse; hybrid architecture; ETL pipeline; ClickHouse; predictive monitoring; Z-score; Service Level Agreement (SLA)

Цифровизация государственного управления и переход к концепции Data-Driven Governance сопровождаются экспоненциальным ростом объемов транзакционных данных. Современная инфраструктура электронного правительства, ядром которой выступают Единый портал государственных и муниципальных услуг (ЕПГУ) и Система межведомственного электронного взаимодействия (СМЭВ), ежедневно генерирует миллионы событий. Каждое обращение гражданина инициирует сложную цепочку межведомственных запросов, статусы которых фиксируются в гетерогенных ведомственных информационных системах (ВИС).

В условиях непрерывного потока телеметрии традиционные подходы к построению систем мониторинга демонстрируют концептуальные

ограничения. Использование классических нормализованных реляционных баз данных (OLTP), ориентированных на обеспечение строгой транзакционной целостности, оказывается неэффективным при решении задач аналитической обработки (OLAP). Выполнение комплексных агрегационных запросов, требующих многоуровневых операций соединения таблиц (JOIN) на массивах из десятков миллионов строк, приводит к критической деградации производительности вычислительных кластеров. В связи с этим актуализируется научно-практическая задача проектирования гибридных хранилищ данных (Data Warehouse, DWH), способных обеспечить высокую скорость потоковой загрузки транзакций, оперативное выполнение предиктивной аналитики и изоляцию нормативно-справочной информации.

Архитектурным фундаментом разрабатываемой системы мониторинга выступает денормализованная многомерная модель данных по схеме «Звезда» (Star Schema). В отличие от классической нормализации (3NF) или схемы «Снежинка», данная структура минимизирует количество связей между сущностями, что является критическим фактором для обеспечения максимальной пропускной способности при сканировании аналитических витрин.

Ключевым этапом инфологического проектирования является фиксация уровня зернистости данных (data granularity). Для обеспечения возможности выявления микросекундных инфраструктурных сбоев шлюзов СМЭВ и локализации «узких мест» административных регламентов выбрана максимальная — транзакционная — зернистость. Строка центральной таблицы фактов фиксирует не агрегированные за сутки показатели ведомства, а каждое атомарное событие жизненного цикла отдельной заявки: смену аналитического статуса, фиксацию технической ошибки или факт превышения регламентного интервала (SLA).

Информационный массив жестко сегментирован на две категории:

- Массив фактов (Log Data): количественные показатели динамики процессов (идентификаторы заявок, временные метки, длительность обработки, флаги нарушений).
- Измерения (Dimensions): контекстуальные аналитические разрезы, формирующие нормативно-справочную информацию (НСИ) — иерархия органов исполнительной власти, классификаторы государственных услуг, типы системных ошибок и производственные календари.

Для реализации спроектированной модели применена гибридная физическая архитектура, физически разделяющая вычислительную нагрузку между двумя специализированными системами управления базами данных: PostgreSQL и ClickHouse.

Реляционный сегмент DWH развернут в СУБД PostgreSQL. Данный контур отвечает за хранение и версионирование нормативно-справочной информации (таблицы измерений `dim_services`, `dim_departments`, `dim_errors`). Использование PostgreSQL обусловлено строгими требованиями к поддержанию свойств ACID (атомарность, согласованность, изолированность, долговечность) и ссылочной целостности при обновлении мастер-данных. Первичные ключи таблиц измерений реализованы через автоинкрементный тип SERIAL, что обеспечивает автоматическое построение компактных B-Tree индексов и ускоряет поиск. Текстовые атрибуты (наименования ведомств, описания ошибок) строго лимитированы типами VARCHAR, что минимизирует паразитный расход дисковой квоты.

Аналитический сегмент, аккумулирующий транзакционные логи, реализован в колоночной СУБД ClickHouse. Центральная таблица фактов (`fact_service_requests`) построена на базе специализированного движка MergeTree. Отказ от строчного формата хранения в пользу колоночного

обеспечивает эффективное применение алгоритмов компрессии (LZ4), многократно снижающих объем данных на физических накопителях (I/O) и ускоряющих операции чтения при расчете агрегатных функций (SUM, AVG, COUNT).

Для предотвращения деградации производительности дисковой подсистемы при экспоненциальном росте базы данных реализовано секционирование (партиционирование) массива по месяцам с использованием выражения `PARTITION BY toYYYYMM(date_id)`. Физическая сортировка данных на диске задана ключом `ORDER BY (service_id, dept_id, date_id)`, что полностью коррелирует с наиболее частотными паттернами фильтрации данных в целевых BI-дашбордах. Параметр гранулярности индекса (`index_granularity`) зафиксирован на уровне 8192 строк, что позволяет удерживать разреженный первичный индекс в оперативной памяти сервера (RAM) в полном объеме.

Интеграция гетерогенных контуров хранилища обеспечивается встроенным механизмом ClickHouse Dictionaries (внешние словари). СУБД ClickHouse осуществляет регламентное чтение таблиц измерений из PostgreSQL и кэширует их в оперативной памяти в виде оптимизированных хэш-таблиц. При поступлении аналитического запроса операция соединения (JOIN) транзакционных фактов с текстовыми справочниками выполняется непосредственно в RAM, исключая накладные расходы на сетевое взаимодействие и транзакционные блокировки в PostgreSQL.

Надежность функционирования системы мониторинга напрямую зависит от отказоустойчивости процессов извлечения, трансформации и загрузки данных (ETL). В качестве инструмента координации потоков данных применен open-source фреймворк Apache Airflow. Логика обработки формализована в виде направленных ациклических графов (DAG) на языке

Python. Скрипты конвейера спроектированы в соответствии с принципом идемпотентности: в случае сетевого сбоя повторный запуск задачи не приводит к дублированию записей в целевых таблицах ClickHouse.

Процесс обработки данных разделен на три технологических этапа:

1) Extract (Извлечение): Реализована инкрементальная стратегия загрузки. Скрипты опрашивают инфраструктуру СМЭВ и операционные базы ведомств, извлекая исключительно новые или измененные пакеты данных за прошедшие сутки. Это радикально снижает нагрузку на продуктивные контуры электронного правительства.

2) Transform (Трансформация): В промежуточной области (Staging Area) сырые данные подвергаются алгоритмической очистке библиотеками Pandas. Производится удаление «технического мусора» (дублирующихся пакетов от повторных сетевых запросов), синтаксическая нормализация временных меток, а также вычисление плановых сроков исполнения услуг. Критически важным этапом трансформации является процесс логического выравнивания (Data Lineage) — маппинг разрозненных текстовых статусов из десятков ВИС в единый числовой аналитический классификатор (status\_id), обеспечивающий сквозную прослеживаемость заявки.

3) Load (Загрузка): Физическая запись данных дифференцирована. Обновление справочников в PostgreSQL происходит одиночными транзакционными вставками. Загрузка логов в ClickHouse осуществляется пакетным методом (Batch Load) блоками по 50–100 тысяч строк. Данный объем блока является оптимальным для архитектуры движка MergeTree, предотвращая излишнюю фрагментацию данных на диске (образование мелких parts) и сохраняя высокую пропускную способность кластера.

Для перехода от описательной статистики к интеллектуальному анализу в структуру ETL-конвейера имплементированы два алгоритма обработки

телеметрии, ориентированных на различные классы государственных сервисов.

Первый алгоритм обеспечивает контроль высоконагруженных синхронных сервисов (например, межведомственных запросов при записи к врачу), где приоритетом является техническая доступность (Uptime) и выявление деградации инфраструктуры. Использование статических порогов времени отклика (статичный таймаут) признано неэффективным ввиду генерации ложных срабатываний в периоды пиковых пользовательских нагрузок. В систему внедрен метод статистического контроля процессов (SPC) на основе динамического вычисления  $Z$ -индекса ( $Z$ -score).

Алгоритм в реальном времени рассчитывает скользящее математическое ожидание  $\mu$  (среднее время отклика) и среднеквадратическое отклонение  $\sigma$  для окна из  $n$  последних транзакций. Каждая новая транзакция с фактическим временем обработки  $t$  оценивается по формуле стандартизированного отклонения:

$$Z = (t - \mu) / \sigma$$

В случае превышения  $Z$ -индексом порогового значения 3.0 (эмпирическое правило «трех сигм») задержка классифицируется не как стохастический сетевой шум, а как структурная аномалия шлюза. Запись автоматически помечается флагом технической ошибки (`is_technical_error = 1`), инициируя инцидент в системе.

Второй алгоритм ориентирован на асинхронные гибридные услуги (например, оформление паспорта) и решает задачу предиктивного выявления рисков нарушения административных регламентов (SLA). Оценка вероятности просрочки осуществляется через вычисление комплексного индекса риска  $I_{risk}$ , который учитывает не только затраченное время  $t_{passed}$  относительно нормативного срока  $t_{sla}$ , но и коэффициент операционной эффективности

конкретного ведомства. Коэффициент рассчитывается как отношение исторического среднего времени выполнения услуги проверяемым подразделением  $t_{historical}$  к среднему региональному показателю  $t_{regional}$ :

$$I_{risk} = (t_{passed} / t_{sla} ) \times (t_{historical} / t_{regional} )$$

Данная математическая модель позволяет классифицировать транзакции по зонам риска (зеленая, желтая, красная) и проактивно сигнализировать в ВІ-систему о необходимости административного вмешательства до момента фактического истечения регламентного срока.

Спроектированная гибридная архитектура хранилища данных на базе связки ClickHouse и PostgreSQL, интегрированная с оркестратором Apache Airflow, формирует технологически суверенный, масштабируемый фундамент для мониторинга процессов электронного правительства. Выделение высоконагруженных операций записи и чтения транзакционных логов в специализированную колоночную СУБД полностью устраняет проблему деградации производительности при работе с массивами класса Big Data.

Внедрение в пайплайны обработки данных математического аппарата на базе алгоритмов динамического Z-индекса и предиктивного расчета рисков SLA трансформирует аналитическую инфраструктуру из пассивного регистратора инцидентов в интеллектуальную систему упреждающего контроля. Результаты проведенного исследования доказывают, что предложенная архитектура обеспечивает минимальный отклик интерфейсов ВІ-дашбордов и предоставляет органам государственной власти объективный цифровой инструментарий для принятия управленческих решений.

### **Список литературы**

1. Кимбалл, Р. Хранилища данных. Руководство по размерному моделированию / Р. Кимбалл, М. Росс. – Москва : Вильямс, 2014. – 712 с.

2. Клеппман, М. Высоконагруженные приложения. Программирование, масштабирование, поддержка / М. Клеппман. – Санкт-Петербург : Питер, 2023. – 638 с. – ISBN 978-5-4461-0512-0.
3. ООН. Электронное правительство 2024 : ускорение цифровой трансформации для устойчивого развития / Департамент по экономическим и социальным вопросам Организации Объединённых Наций. – Нью-Йорк : ООН, 2024. – URL: <https://desapublications.un.org/sites/default/files/publications/2025-01/E-Government%20Survey%202024%20RUS-compressed.pdf> (дата обращения: 10.05.2026).
4. Montgomery, D. C. Introduction to Statistical Quality Control / D. C. Montgomery. – 8th ed. – Hoboken : John Wiley & Sons, 2019. – 776 p.
5. Ruiter, J. de. Data Pipelines with Apache Airflow. 2nd Edition. Orchestration for Data and AI / J. de Ruiter, I. Cabral, K. Geusebroek, D. van der Ende, B. Harenslak. – Shelter Island : Manning Publications, 2026. – 600 p. – ISBN 978-1-633436-37-4.
6. Apache Airflow Documentation [Электронный ресурс] // Apache Software Foundation. – URL: <https://airflow.apache.org/docs/> (дата обращения: 22.05.2026).
7. ClickHouse Documentation [Электронный ресурс] // ClickHouse Inc. – URL: <https://clickhouse.com/docs/> (дата обращения: 14.05.2026).
8. pandas Documentation [Электронный ресурс] // The pandas Development Team. – URL: <https://pandas.pydata.org/docs/> (дата обращения: 20.05.2026).
9. PostgreSQL Documentation [Электронный ресурс] // The PostgreSQL Global Development Group. – URL: <https://www.postgresql.org/docs/> (дата обращения: 17.05.2026).

## References

1. Kimball R., Ross M. Data Warehouses: A Guide to Dimensional Modeling [In Russian]. Moscow: Williams, 2014. 712 p.

2. Kleppmann M. Designing Data-Intensive Applications: Programming, Scaling, and Supporting [In Russian]. St. Petersburg: Piter, 2023. 638 p. ISBN 978-5-4461-0512-0.
3. United Nations Department of Economic and Social Affairs. United Nations E-Government Survey 2024: Accelerating Digital Transformation for Sustainable Development [In Russian]. New York: United Nations, 2024. Available at:  
<https://desapublications.un.org/sites/default/files/publications/2025-01/E-Government%20Survey%202024%20RUS-compressed.pdf> (accessed May 10, 2026).
4. Montgomery D. C. Introduction to Statistical Quality Control. 8th ed. Hoboken: John Wiley & Sons, 2019. 776 p.
5. de Ruiter J., Cabral I., Geusebroek K., van der Ende D., Harenslak B. Data Pipelines with Apache Airflow. 2nd Edition: Orchestration for Data and AI. Shelter Island, NY: Manning Publications, 2026. ISBN 978-1-633436-37-4.
6. Apache Software Foundation. Apache Airflow Documentation. Available at: <https://airflow.apache.org/docs/> (accessed May 22, 2026).
7. ClickHouse Inc. ClickHouse Documentation. Available at: <https://clickhouse.com/docs/> (accessed May 14, 2026).
8. The pandas Development Team. pandas Documentation. Available at: <https://pandas.pydata.org/docs/> (accessed May 20, 2026).
9. The PostgreSQL Global Development Group. PostgreSQL Documentation. Available at: <https://www.postgresql.org/docs/> (accessed May 17, 2026).