АРХІТЕКТУРНІ ПІДХОДИ ДО ЯКОСТІ ТА БЕЗПЕКИ ДАНИХ В АНАЛІТИЧНИХ ПЛАТФОРМАХ ОХОРОНИ ЗДОРОВ'Я

Печій Р.Р. Шубін І.Ю.

Кафедра інфокомунікаційної інженерії ім. В.В. Поповського,
Харківський національний університет радіоелектроніки,
Україна

E-mail: rostyslav.pechii@nure.ua,
igor.shubin@nure.ua

Abstract

This article examines architectural approaches to ensuring data quality and cybersecurity specifically within healthcare analytical platforms designed for Revenue Cycle Management (RCM). It addresses the primary security threat in healthcare—the human element—and its implications for big data systems that ingest heterogeneous, semi-structured data. The paper argues that traditional security perimeters are insufficient when social engineering and insider threats are the main attack vectors. Consequently, robust, data-centric security controls must be embedded within the analytical platform itself. It then explores the architectural challenges of ensuring data integrity and quality when classic normalization techniques are not viable due to the volume and variety of incoming data (charges, payments, denials). The study proposes a data governance and security framework based on resilient ETL/ELT pipelines, master data management (MDM), and the enforcement of security policies like TDE and RLS to meet HIPAA's stringent requirements in a big data context..

Цифровізація системи охорони здоров'я перетворила медичні дані на найцінніший актив галузі, але водночас і на її головний епіцентр ризику. Згідно зі звітом IBM «Cost of a Data Breach Report 2023», сектор охорони здоров'я вже 13-й рік поспіль зазнає найвищих фінансових збитків від витоків даних, що в середньому складає $10.93 мільйона за інцидент. Ця тенденція пояснюється високою вартістю електронної захищеної медичної інформації (ePHI) на чорному ринку, зумовленою її довготривалим характером та широкими можливостями для шахрайства.

Аналітичні платформи для управління доходами (Revenue Cycle Management, RCM) стали зовнішім аналітичним центром сучасних медичних установ. Вони агрегують колосальні обсяги фінансових та операційних даних – мільйони транзакцій щодня по рахунках (Accounts), платежах (Payments), нарахуваннях (Charges) та відмовах у виплатах (Denials) – для виявлення фінансових аномалій та оптимізації доходів. Концентрація таких даних створює надзвичайно цінний актив, що, своєю чергою, робить ці платформи головною ціллю для кібератак.

Центральною проблемою при проєктуванні таких систем є дотримання вимог Акту про мобільність та підзвітність медичного страхування (HIPAA). Незважаючи на те, що дані переважно фінансові, вони нерозривно пов'язані з пацієнтами та наданими їм послугами, а отже, класифікуються як захищена медична інформація (ePHI). Правило безпеки HIPAA вимагає від організацій забезпечення трьох фундаментальних властивостей даних: конфіденційності, цілісності та доступності. Порушення цих вимог веде до колосальних фінансових та репутаційних втрат.

Аналіз сучасних інцидентів безпеки в охороні здоров'я показує, що переважна більшість витоків даних відбувається не через складні технічні зломи інфраструктури, а через людський фактор. Соціальна інженерія, фішингові атаки, спрямовані на крадіжку облікових даних співробітників, та ненавмисні помилки персоналу є головним вектором атак. Зловмисник, отримавши доступ до системи від імені співробітника, може обійти периметри безпеки, такі як міжмережеві фаєрволи. Це означає,

що безпека не може більше покладатися лише на зовнішній захист; вона має бути інтегрована безпосередньо в архітектуру самої аналітичної платформи.

Цей виклик ускладнюється природою даних, з якими працюють RCM-системи. Дані надходять з десятків, а іноді й сотень різних джерел: білінгові системи, електронні медичні картки (EHR), дані страхових компаній. Ці дані є гетерогенними, часто неструктурованими або напівструктурованими, містять помилки, пропуски та розбіжності у форматах. Вони аналізуються «по факту» з того, що є, оскільки головна цінність аналітики полягає у виявленні патернів саме в цьому сирому потоці інформації.

У таких умовах класичний підхід до проєктування баз даних, що базується на нормалізації, стає нежиттєздатним. Спроба привести весь цей хаотичний потік даних до суворої, нормалізованої схеми перед завантаженням призвела б до двох негативних наслідків. По-перше, процес трансформації та завантаження (ETL) був би надзвичайно повільним, що унеможливило б аналіз даних у режимі, близькому до реального часу. По-друге, жорстка схема відкинула б або спотворила б «нестандартні» дані, які часто містять найважливішу інформацію про аномалії в процесах. Таким чином, для забезпечення продуктивності та повноти аналізу, RCM-платформи змушені працювати з денормалізованими або слабко структурованими даними.

Це створює ключову архітектурну дилему: як забезпечити цілісність та якість даних, як того вимагає HIPAA, в середовищі, де відсутній головний інструмент для цього – нормалізація, і як захистити ці дані від несанкціонованого доступу, коли головною загрозою є легітимний користувач, чиї облікові дані були скомпрометовані?

Оскільки головною загрозою є людина (через фішинг або помилки), архітектура безпеки має бути зосереджена на мінімізації збитків від скомпрометованого облікового запису. Це досягається за допомогою вбудованих, даних-орієнтованих механізмів СУБД:

- Прозоре шифрування даних (TDE) є обов'язковим базовим рівнем, що захищає дані «у спокої» від фізичної крадіжки носіїв, але воно безсиле проти атак зсередини системи.
- Рольова модель доступу (RBAC) та Безпека на рівні рядків (RLS) є головними інструментами протидії. За допомогою RBAC визначаються ролі (наприклад, «Фінансовий аналітик») з чітко обмеженими правами. RLS йде далі і дозволяє обмежити видимість даних всередині однієї таблиці.

Проєктування аналітичних платформ для управління доходами в охороні здоров'я вимагає відходу від класичних архітектурних шаблонів. Основна проблема полягає в тому, що традиційні методи забезпечення цілісності даних, зокрема нормалізація, є нежиттєздатними в середовищі великих даних через їх катастрофічний вплив на продуктивність аналітичних запитів, що прямо суперечить вимогам доступності даних згідно з HIPAA.

Оскільки класична нормалізація є неприйнятною для аналітичних RCM-платформ, необхідно застосовувати архітектурні підходи, розроблені спеціально для сховищ даних та бізнес-аналітики. Головним таким підходом є вимірне моделювання (dimensional modeling), найпопулярнішою реалізацією якого є схема «зірка» (star schema). Ця модель передбачає контрольовану денормалізацію, де дані організовані навколо центральної «таблиці фактів» (наприклад, транзакції по нарахуваннях), яка містить кількісні показники та ключі до «таблиць вимірів» (довідників пацієнтів, клінік, страхових компаній).

У такій архітектурі відповідальність за забезпечення якості та цілісності даних переноситься з рівня структури бази даних на рівень процесів завантаження даних – ETL/ELT (Extract, Transform, Load / Extract, Load, Transform). Саме в цих процесах сирі, гетерогенні дані з різних джерел проходять етапи очищення, стандартизації, валідації та збагачення перед тим, як потрапити в аналітичне сховище. Порівняльний аналіз ETL-процесів для гіпотетичної нормалізованої та реалістичної денормалізованої схеми наочно демонструє переваги останньої для великих даних.

Таблиця 1. Порівняльний аналіз ETL-процесів для аналітичних схем в MS SQL Server

Характеристика	Нормалізована схема (3NF)	Денормалізована схема ("Зірка")
Кількість таблиць для завантаження (на 1 вхідний запис про нарахування)	5-8 (факти, довідники, але додані всі обмеження і зовнішні ключі.)	5-8 (основна таблиця фактів різні виміри, але немає зовнішніх ключв)
Складність логіки завантаження	Висока (послідовний запис у довідники, перевірка FOREIGN KEY на кожному кроці)	Помірна (пошук ключів у таблицях вимірів, один масовий запис у таблицю фактів)
Час завантаження пакету даних (1 млн. транзакцій нарахувань)	~55 хвилин	~7 хвилин
Продуктивність аналітичного запиту (звіт по відмовах з JOIN 5 таблиць)	~25-30 хвилин	~3-5 секунд
Обсяг дискового простору	~1.5 ТБ	~2.2 ТБ

Як видно з таблиці, спроба завантажити великий обсяг даних у суворо нормалізовану схему є надзвичайно повільною через необхідність виконувати безліч операцій для перевірки цілісності та запису в різні таблиці. Час завантаження стає вузьким місцем, що унеможливлює оперативний аналіз.

Найбільш критичною є різниця у продуктивності аналітичних запитів: хвилини у нормалізованій схемі проти секунд у денормалізованій. Це відбувається тому, що в схемі «зірка» ресурсоємні операції JOIN фактично були виконані один раз на етапі ETL, а не виконуються заново при кожному запиті. Ми свідомо жертвуємо ефективністю зберігання (денормалізована схема займає більше місця через надлишковість) заради радикального прискорення доступу до даних, що є ключовою вимогою для аналітичних систем і прямо впливає на доступність даних згідно з HIPAA.

Запропонований підхід базується на визнанні цього конфлікту та перенесенні відповідальності за якість даних з рівня структури БД на рівень процесів ETL/ELT. Саме на цьому етапі відбувається очищення, стандартизація та валідація даних, що дозволяє завантажувати їх у контрольовано денормалізовані сховища, оптимізовані для швидкого читання, такі як схема «зірка». Ця архітектура свідомо жертвує ефективністю зберігання заради радикального підвищення продуктивності аналітики.

Водночас, оскільки головною загрозою безпеці є людський фактор та скомпрометовані облікові дані, архітектура захисту має бути зосереджена на мінімізації потенційних збитків. Це досягається шляхом впровадження глибоко інтегрованих, даних-орієнтованих механізмів безпеки, таких як Рольова модель доступу (RBAC) та Безпека на рівні рядків (RLS). Ці інструменти дозволяють сегментувати доступ до даних на логічному рівні, гарантуючи, що навіть легітимний користувач (або зловмисник під його обліковим записом) матиме доступ лише до мінімально необхідного набору даних. Таким чином, створюється комплексна, ешелонована система, що збалансовує вимоги до продуктивності, якості та безпеки в складному регуляторному середовищі HIPAA.

Література

1. BM Security (2023). Cost of a Data Breach Report 2023. [Online]. Available: https://www.ibm.com/reports/data-breach.

2. Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (3rd ed.). Wiley.

3. U.S. Department of Health & Human Services. (2013). The Security Rule. Health Insurance Portability and Accountability Act of 1996 (HIPAA). [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/security/index.html.

4. Verizon (2023). 2023 Data Breach Investigations Report (DBIR). [Online]. Available: https://www.verizon.com/business/resources/reports/dbir/.

5. Microsoft Corporation (2022). Row-Level Security in SQL Server. SQL Server Technical Documentation. [Online]. Available: https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security.

6. Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. Communications of the ACM, 13(6), pp. 377–387.