# БЕЗОПАСНЫЕ ОБЛАЧНЫЕ ХРАНИЛИЩА БИОМЕДИЦИНСКИХ ДАННЫХ

Хлебанова К. О.
Факультет компьютерного проектирования,
Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: ksenia.hlebanova@gmail.com

*В последние годы объем биомедицинских данных, генерируемых клиническими системами, геномными исследованиями и носимыми устройствами, стремительно возрастает. Это требует разработки архитектур облачных хранилищ, обеспечивающих высокую производительность, масштабируемость и защиту персональных медицинских данных. В работе рассматриваются современные подходы к организации хранения биомедицинских данных в облаке, включая распределённые системы, микросервисные архитектуры и механизмы обеспечения конфиденциальности и целостности информации.*

## Введение

Биомедицинские данные характеризуются высокой гетерогенностью (большие бинарные объекты – медицинские изображения, большие файлы геномики; структурированные EHR/метаданные; потоковые сенсорные данные), строгими требованиями к конфиденциальности (PHI, GDPR/HIPAA), потребностью в аналитике в реальном времени и архивном хранении [1-2]. Проектируемая архитектура должна обеспечивать:

- масштабируемость по объёму и запросам;
- различную производительность для горячих и холодных данных;
- контроль доступа, шифрование и доказуемый аудит;
- поддержку стандартов обмена (FHIR, DICOM, HL7) и возможность приватных вычислений (secure enclave, TEEs, SMC/HE).

Далее рассматриваются два подхода, реалистичных для облачной среды.

## I. Централизованный облачный Data Lake

Централизованный облачный *Data Lake* представляет собой архитектуру, где объектное хранилище (S3-подобного типа) служит единым источником данных – DICOM, геномных (FASTQ/BAM) и электронных медицинских записей (EHR). Метаданные и индексы хранятся в NoSQL или реляционной базе, связанной с FHIR-ресурсами. Поток данных проходит через слой преобразования (ETL/ELT), нормализующий их в аналитические форматы Parquet/ORC.

Аналитический слой (BigQuery, Athena, Spark) обеспечивает масштабируемую пакетную аналитику. Безопасность реализуется с помощью шифрования (at-rest, in-transit), IAM, KMS и аудита. Для оптимизации затрат используется многоуровневое хранение (hot–cold) с автоматическим tiering. DICOM-объекты индексируются через метаданные и миниатюры, геномные данные преобразуются в табличный формат, EHR – в FHIR-структуры [2]. Время отклика системы можно аппроксимировать выражением:

$$L = L_{\text{net}} + L_{\text{meta}} + L_{\text{io}}, \quad L_{\text{io}} \approx \frac{S}{B},$$

где $L_{\text{net}}$ – сетевые задержки при передаче данных, $L_{\text{meta}}$ – время поиска и обработки метаданных, $L_{\text{io}}$ – задержка при чтении объектов из хранилища, зависящая от их размера $S$ и пропускной способности канала $B$.

Надёжность системы определяется вероятностью доступности данных при $k$-кратной репликации:

$$A = 1 - (1 - a)^k,$$

где $a$ – индивидуальная надёжность отдельного узла хранения, а $A$ – совокупная доступность объекта. Увеличение числа копий данных повышает отказоустойчивость, но увеличивает издержки хранения.

Стоимость хранения оценивается как:

$$Cost = V_{\text{hot}} c_{\text{hot}} + V_{\text{cold}} c_{\text{cold}} + T_{\text{egress}} c_{\text{egress}},$$

где $V_{\text{hot}}$ и $V_{\text{cold}}$ – объёмы данных в горячем и холодном уровнях хранения, $c_{\text{hot}}$ и $c_{\text{cold}}$ – их удельные стоимости, $T_{\text{egress}}$ – объём исходящего трафика, а $c_{\text{egress}}$ – стоимость передачи. Данная модель позволяет балансировать между производительностью и затратами при больших объёмах медицинских данных.

Преимущества: централизованное управление, готовность к аналитике, эффективное хранение крупных файлов. Недостатки: единая точка отказа, риск утечки PHI, возможные сетевые задержки при высокой нагрузке. Подход эффективен для крупных исследовательских и национальных платформ с централизованным контролем доступа.

## II. Федеративная архитектура с безопасными вычислениями

Федеративная (*Data Mesh / Federated*) архитектура строится на распределённой модели, где каждое доменное хранилище (клиника или

лаборатория) сохраняет собственные данные локально или в изолированном облачном пространстве. Глобальный каталог метаданных содержит только описания и минимальные индексы, а взаимодействие между доменами осуществляется через контракты данных и единые API-интерфейсы (FHIR). Центральная оркестрационная платформа маршрутизирует запросы, выполняет авторизацию и обеспечивает виртуализацию данных, не требуя их физического перемещения.

Вычисления над конфиденциальными данными выполняются в доверенных средах (*secure enclaves*) или с применением методов защищённых вычислений – многопартийных протоколов (MPC) и гомоморфного шифрования (HE). Для кросс-доменных исследований применяются федеративное обучение и приватизация данных по модели дифференциальной приватности.

Время выполнения распределённого запроса описывается выражением:

$$T = \max_{i \in [1..n]} (T_{local,i}) + T_{agg} + T_{net},$$

где $T_{local,i}$ – время обработки на домене $i$, $T_{agg}$ – агрегация результатов, $T_{net}$ – сетевые накладные расходы. Стоимость распределённого запроса определяется как:

$$Cost_{\text{query}} = \sum_i cost_{\text{compute},i} + cost_{\text{orchestration}},$$

где $cost_{\text{compute},i}$ – вычислительная стоимость обработки на каждом домене, а $cost_{\text{orchestration}}$ – накладные расходы на координацию, маршрутизацию и объединение результатов между участниками федерации. Применение $\varepsilon$-дифференциальной приватности позволяет ограничить вероятность раскрытия персональных данных при агрегации.

Основными преимуществами подхода являются высокая степень защиты данных, соответствие юридическим требованиям (данные не покидают юрисдикцию) и распределённый контроль. Недостатками выступают сложность согласования форматов и политик доступа, повышенные задержки при агрегации и более высокие операционные затраты. Такой подход предпочтителен при организации межведомственных или международных биомедицинских систем, где конфиденциальность превалирует над скоростью обработки.

## III. Сравнительный анализ и вывод

Централизованный *Data Lake* обеспечивает высокую производительность и удобство аналитической обработки за счёт унифицированного хранилища и зрелых облачных инструментов. Однако он создаёт единую точку отказа и повышенные риски компрометации конфиденциальных медицинских данных. Федеративная архитектура, напротив, распределяет ответственность между доменами, снижает вероятность утечек и лучше соответствует нормативным требованиям (HIPAA, GDPR), но требует сложной оркестрации и имеет большие сетевые задержки.

Сравнение по ключевым параметрам представлено в табл. III.

Таблица 1 – Сравнение архитектурных подходов

| **Критерий** | **Централизованный Data Lake** | **Федеративная архитектура** |
|---|---|---|
| Производительность | Высокая, оптимизированные OLAP-запросы | Зависит от числа доменов и сети |
| Безопасность и приватность | Централизованное управление, высокий риск утечек | Распределённый контроль, минимизация PHI-передачи |
| Масштабируемость | Простое горизонтальное масштабирование | Ограничено политиками и совместимостью |
| Соответствие регуляторным требованиям | Ограничено при кросс-юрисдикционном обмене | Полное соответствие локальному законодательству |
| Операционные затраты | Ниже при большом объёме данных | Выше из-за оркестрации и синхронизации |

Оптимальным направлением развития является гибридный подход, объединяющий сильные стороны обоих решений: использование централизованного *Data Lake* для обезличенных или агрегированных данных и федеративных механизмов для чувствительных данных. Такая гибридная архитектура обеспечивает баланс между безопасностью, масштабируемостью и аналитической эффективностью, что делает её наиболее перспективной для будущих облачных платформ хранения и анализа биомедицинской информации.

## IV. Список литературы


1. Tahir, A.; Rehman, A.; Iqbal, S. et al. A Systematic Review on Cloud Storage Mechanisms in Healthcare / A. Tahir, A. Rehman, S. Iqbal // Sensors. – 2020. – Vol. 20, № 18. – P. 5392.
2. Hashem, I. A. T.; Yaqoob, I.; Anuar, N. B. et al. The rise of "big data" on cloud computing: Review and open research issues / I. A. T. Hashem, I. Yaqoob, N. B. Anuar // Information Systems. – 2015. – Vol. 47. – P. 98–115.
3. Manco, C.; et al. HEALER: A Data Lake Architecture for Healthcare / C. Manco, et al. // CEUR Workshop Proceedings, Vol. 3379. – 2023. – P. 602
4. Simhadri, S. Y.; Deshpande, L. K. Evaluating Data Mesh and Traditional Data Architectures: Implications for Governance, Scalability, Cost, and Enterprise Adoption / S. Y. Simhadri, L. K. Deshpande // Frontiers in Health Informatics. – 2024. – Vol. 13, № 8. – P. 5095–5106.