

---

# Потоковые базы данных

*Объединение пакетной и потоковой обработки*

*Хьюберт Дюлей и Ральф М. Дебусманн*

**SPRINT**  
book 2025

Выпущено  
при поддержке  
**КРОК**

---

# Краткое содержание

|  |     |
|--|-----|
| Введение .....   | 12  |
| Предисловие .....  | 15  |
| От издательства .....  | 18  |
| <b>Глава 1.</b> Основы потоковой обработки .....                       | 20  |
| <b>Глава 2.</b> Платформы обработки потоков .....                      | 40  |
| <b>Глава 3.</b> Предоставление данных в режиме реального времени ..... | 61  |
| <b>Глава 4.</b> Материализованные представления.....                   | 81  |
| <b>Глава 5.</b> Поточковые базы данных .....                           | 103 |
| <b>Глава 6.</b> Согласованность .....                                  | 124 |
| <b>Глава 7.</b> Появление других гибридных систем данных .....         | 158 |
| <b>Глава 8.</b> Zero-ETL или Near-Zero-ETL.....                        | 174 |
| <b>Глава 9.</b> Плоскость потоковой обработки .....                    | 197 |
| <b>Глава 10.</b> Модели развертывания.....                             | 215 |
| <b>Глава 11.</b> Будущее данных в режиме реального времени .....       | 233 |
| Об авторах .....   | 267 |
| Иллюстрация на обложке.....  | 268 |
| Алфавитный указатель .....   | 269 |

---

# Оглавление

|   |           |
|---|-----------|
| Введение .....  | 12        |
| Предисловие .....   | 15        |
| Условные обозначения.....   | 15        |
| Использование примеров кода .....   | 16        |
| Благодарности от Хьюберта.....  | 17        |
| Благодарности от Ральфа.....  | 17        |
| От издательства .....   | 18        |
| О научном редакторе русскоязычного издания .....                                      | 18        |
| Словарь терминов.....   | 18        |
| <b>Глава 1. Основы потоковой обработки.....</b>                                       | <b>20</b> |
| Выворачиваем базу данных наизнанку .....  | 22        |
| Вынос функциональности базы данных за ее пределы .....                                | 22        |
| Механизм упреждающей записи.....  | 23        |
| Платформы потоковой обработки.....  | 25        |
| Материализованные представления .....   | 29        |
| Пример практического применения: анализ потоков цифровых следов<br>пользователей..... | 30        |
| Изучение транзакций и событий.....  | 31        |
| Предметно-ориентированное проектирование .....  | 31        |
| Обогащение контекста .....  | 33        |
| Захват изменения данных .....   | 33        |
| Коннекторы .....  | 35        |
| Промежуточное программное обеспечение коннекторов .....                               | 36        |
| Встроенные коннекторы.....  | 37        |
| Специально разработанные коннекторы .....   | 37        |
| Резюме .....  | 39        |

|  |    |
|--|----|
| <b>Глава 2. Платформы обработки потоков</b> .....                                | 40 |
| Преобразования с состоянием.....   | 42 |
| Конвейеры обработки данных.....  | 45 |
| Ограничения ELT .....  | 48 |
| Обработка потоков с помощью ELT .....  | 49 |
| Потоковые процессоры .....   | 50 |
| Популярные потоковые процессоры.....   | 51 |
| Новые потоковые процессоры.....  | 51 |
| Эмуляция материализованных представлений в Apache Spark.....                     | 53 |
| Два типа потоков.....  | 53 |
| Поток с добавлением .....  | 55 |
| Изменение данных с Debezium.....   | 56 |
| Материализованные представления .....  | 57 |
| Резюме .....   | 60 |
| <b>Глава 3. Предоставление данных в режиме реального времени</b> .....           | 61 |
| Ожидаемый результат работы в режиме реального времени.....                       | 62 |
| Выбор аналитического хранилища данных.....                                       | 63 |
| Топик в качестве источника .....   | 65 |
| Преобразования при поглощении .....  | 66 |
| OLTP и OLAP .....  | 68 |
| ACID .....   | 69 |
| Строчная и столбцовая оптимизации .....  | 70 |
| Запросы в секунду и конкурентность.....  | 71 |
| Индексирование.....  | 72 |
| Предоставление результатов анализа .....   | 76 |
| Синхронные запросы.....  | 76 |
| Асинхронные запросы .....  | 77 |
| Запросы push и pull.....   | 78 |
| Резюме .....   | 80 |
| <b>Глава 4. Материализованные представления</b> .....                            | 81 |
| Представления, материализованные представления и инкрементные<br>обновления..... | 82 |
| Захват изменения данных .....  | 84 |
| Запросы push и pull.....   | 86 |

|  |            |
|--|------------|
| CDC и Upsert .....   | 91         |
| Объединение потоков .....  | 94         |
| Apache Calcite .....   | 95         |
| Пример использования цифрового следа .....   | 99         |
| Резюме .....   | 101        |
| <b>Глава 5. Поточковые базы данных .....</b>   | <b>103</b> |
| Определение потоковой базы данных .....  | 104        |
| Потоковая база данных на основе столбцов .....   | 107        |
| Потоковая база данных на основе строк .....  | 108        |
| Граничные базы данных, подобные потоковым .....  | 111        |
| Выразительность SQL .....  | 111        |
| Возможности отладки потоковой обработки .....  | 114        |
| Преимущества отладки в потоковых базах данных .....  | 115        |
| SQL не панацея .....   | 115        |
| Реализации потоковых баз данных .....  | 116        |
| Архитектура потоковой базы данных .....  | 117        |
| Конвейеры ELT с потоковыми базами данных .....   | 121        |
| Резюме .....   | 122        |
| <b>Глава 6. Согласованность .....</b>  | <b>124</b> |
| Воображаемый пример .....  | 125        |
| Транзакции .....   | 126        |
| Анализ транзакций .....  | 127        |
| Сравнение согласованности между системами обработки потоков .....  | 128        |
| Flink SQL .....  | 128        |
| ksqlDB .....   | 132        |
| Proton (Timeplus) .....  | 135        |
| RisingWave .....   | 138        |
| Materialize .....  | 140        |
| Pathway .....  | 142        |
| Выход за рамки согласованности в конечном счете .....  | 145        |
| Почему согласованные в конечном счете потоковые процессоры<br>не справляются с рассмотренным примером? ..... | 145        |
| Как внутренне согласованные системы обработки потоков<br>выполняют рассматриваемый пример? .....             | 149        |

|  |            |
|--|------------|
| Как можно исправить системы обработки потоков с согласованностью в конечном счете, чтобы выполнить рассматриваемый пример? ..... | 152        |
| Согласованность и задержка .....   | 156        |
| Резюме .....   | 157        |
| <b>Глава 7. Появление других гибридных систем данных .....</b>   | <b>158</b> |
| Плоскости данных .....   | 159        |
| Гибридная транзакционно-аналитическая база данных .....  | 161        |
| Другие гибридные базы данных .....   | 165        |
| Мотивы создания гибридных систем .....   | 165        |
| Влияние PostgreSQL на гибридные базы данных .....  | 167        |
| Выполнение аналитики вблизи границы плоскостей данных .....  | 168        |
| Гибридные базы данных нового поколения .....   | 169        |
| Потоковые БД OLTP нового поколения .....   | 170        |
| Потоковые БД RTOLAP нового поколения .....   | 172        |
| Базы данных HTAP нового поколения .....  | 172        |
| Резюме .....   | 173        |
| <b>Глава 8. Zero-ETL или Near-Zero-ETL .....</b>   | <b>174</b> |
| Модель ETL .....   | 174        |
| Zero-ETL .....   | 175        |
| Near-zero-ETL .....  | 178        |
| PeerDB .....   | 178        |
| Proton .....   | 180        |
| Встроенные OLAP .....  | 181        |
| Гравитация и репликация данных .....   | 186        |
| Сокращение аналитических данных .....  | 186        |
| Лямбда-архитектура .....   | 187        |
| Гибридные таблицы Apache Pinot .....   | 189        |
| Конфигурации конвейеров .....  | 194        |
| Резюме .....   | 196        |
| <b>Глава 9. Плоскость потоковой обработки .....</b>  | <b>197</b> |
| Гравитация данных .....  | 198        |
| Компоненты потоковой плоскости .....   | 200        |
| Инфраструктура потоковой плоскости .....   | 202        |

|  |            |
|--|------------|
| Операционная аналитика .....   | 203        |
| Сетка данных .....   | 206        |
| Столпы сетки данных .....  | 207        |
| Трудности применения сетки данных .....                                | 209        |
| Потоковая сетка данных с потоковой плоскостью и потоковыми БД .....    | 210        |
| Локальность данных .....   | 211        |
| Репликация данных .....  | 212        |
| Резюме .....   | 214        |
| <b>Глава 10. Модели развертывания .....</b>                            | <b>215</b> |
| Согласованная потоковая база данных .....                              | 216        |
| Согласованный потоковый процессор и RTOLAP .....                       | 218        |
| Потоковая БД OLAP с согласованностью в конечном счете .....            | 219        |
| Потоковый процессор с согласованностью в конечном счете и RTOLAP ..... | 220        |
| Потоковый процессор с согласованностью в конечном счете и HTAP .....   | 221        |
| ksqlDB .....   | 222        |
| Инкрементное обслуживание представлений .....                          | 223        |
| Обертка сторонних данных Postgres Multicorn .....                      | 224        |
| Когда следует использовать потоковые процессоры на основе кода .....   | 225        |
| Когда следует использовать технологии Lakehouse/Streamhouse .....      | 225        |
| Технологии кэширования .....   | 226        |
| Где выполнять обработку и запросы в общем случае .....                 | 227        |
| Четыре вопроса «Где?» .....  | 227        |
| Аналитический сценарий использования .....                             | 228        |
| Последствия .....  | 230        |
| Резюме .....   | 232        |
| <b>Глава 11. Будущее данных в режиме реального времени .....</b>       | <b>233</b> |
| Слияние плоскостей данных .....  | 234        |
| Графовые базы данных .....   | 235        |
| Memgraph .....   | 236        |
| thatDot/Quine .....  | 237        |
| Векторные базы данных .....  | 240        |
| Milvus 2.x — потоковая передача как основа .....                       | 241        |
| Базы данных RTOLAP — добавление векторного поиска .....                | 243        |

|  |     |
|--|-----|
| Инкрементное обслуживание представлений (IVM) .....            | 244 |
| pg_ivm .....   | 245 |
| Hydra .....  | 245 |
| Epsio .....  | 246 |
| Feldera .....  | 247 |
| PeerDB .....   | 248 |
| Обертывание данных и Postgres Multicorn .....                  | 250 |
| Классические базы данных .....                                 | 253 |
| Хранилища данных .....   | 255 |
| BigQuery .....   | 255 |
| Redshift .....   | 256 |
| Snowflake .....  | 257 |
| Lakehouse .....  | 259 |
| Delta Lake .....   | 260 |
| Apache Paimon .....  | 261 |
| Apache Iceberg .....   | 262 |
| Apache Hudi .....  | 263 |
| OneTable или XTable .....                                      | 264 |
| Взаимоотношения потоковой обработки и хранилищ Lakehouse ..... | 264 |
| Резюме .....   | 266 |
| Об авторах .....   | 267 |
| Иллюстрация на обложке .....                                   | 268 |
| Алфавитный указатель .....                                     | 269 |