
Оглавление

От издательства	13
О научном редакторе русского издания	13
Отзывы о книге «Delta Lake. Современные архитектуры озер данных»	14
Предисловие Майкла Армбруста	16
Предисловие Доминика Брезински	17
Введение	18
Для кого предназначена эта книга	18
Как устроена эта книга	19
Условные обозначения в книге	21
Использование исходного кода примеров	22
Благодарности	23
Дэнни	24
Тристен	24
Скотт	24
Прашант	25
Глава 1. Введение в технологию Delta Lake для озер-хранилищ данных	26
Генезис Delta Lake	26
Хранилища данных, озера данных и озера-хранилища данных	26
От проекта Tahoe к Delta Lake: первые годы месяцы	30
Что такое Delta Lake?	31
Распространенные варианты применения	32
Основные функции	33
Структура таблиц Delta Lake	35
Протокол транзакций Delta	37
Знакомство с журналом транзакций Delta Lake на уровне файлов	38
Единый источник истины	38
Связь между метаданными и данными	39
Рассмотрение файлов и данных в рамках многоверсионного управления конкурентностью (MVCC)	39
Обзор взаимодействия между метаданными и данными	40
Table Features	42

Ядро Delta (Delta Kernel)	44
Delta UniForm	46
Заключение	47
Глава 2. Установка Delta Lake	48
Docker-образ для Delta Lake	48
Delta Lake для Python	50
Оболочка PySpark	51
Блокнот JupyterLab (JupyterLab Notebook)	52
Оболочка Scala	52
Rust API для Delta.....	53
ROAPI (Read-Only API)	54
Нативные библиотеки Delta Lake	55
Доступно несколько привязок	56
Установка пакета Delta Lake для Python	56
Apache Spark с Delta Lake	56
Настройка Delta Lake с Apache Spark	57
Обязательное условие: настройка Java.....	57
Настройка интерактивной оболочки	58
Декларативный API в PySpark	60
Версия Databricks Community Edition	60
Создание кластера с помощью среды выполнения Databricks	60
Импорт блокнотов.....	62
Привязка к блокноту	63
Заключение	64
Глава 3. Основные операции Delta Lake	65
Создание (Create)	66
Создание таблицы Delta Lake	67
Загрузка данных в таблицу Delta Lake	68
Журнал транзакций.....	71
Чтение (Read)	72
Запрос данных из таблицы Delta Lake	72
Чтение данных посредством функции Time Travel.....	73
Обновление (Update)	75
Удаление (Delete)	75
Удаление данных из таблицы Delta Lake	76
Перезапись данных в таблице Delta Lake	77
Слияние (Merge).....	79

Другие полезные операции.....	81
Преобразования из формата Parquet.....	82
Метаданные и история в Delta Lake	83
Заключение	84
Глава 4. Погружение в экосистему Delta Lake	85
Коннекторы	86
Apache Flink	87
Коннектор Flink DataStream	87
Установка коннектора.....	88
API DeltaSource	88
API DeltaSink	92
Полный сквозной пример (End-to-End Example).....	95
Коннектор Kafka Delta Ingest.....	97
Установка Rust	99
Сборка проекта.....	99
Запуск процесса загрузки данных.....	100
Экосистема Trino	102
Начало работы.....	102
Настройка и использование коннектора Trino	106
Команда Show Catalogs (Вывод списка каталогов)	106
Создание схемы	107
Команда Show Schemas (Вывод списка схем)	107
Работа с таблицами	108
Операции с таблицами	111
Заклучение	115
Глава 5. Сопровождение систем на базе Delta Lake	116
Свойства таблиц Delta Lake.....	116
Перечень свойств таблицы Delta Lake	117
Создание пустой таблицы со свойствами.....	119
Заполнение таблицы	119
Эволюция схемы таблицы	121
Добавление или изменение свойств таблицы	123
Удаление свойств таблицы	124
Оптимизация таблиц Delta Lake.....	126
Проблема больших таблиц и мелких файлов	126
Команды OPTIMIZE для решения проблемы маленьких файлов	129

Тонкая настройка и управление таблицами	131
Партиционирование таблиц.....	131
Определение разделов при создании таблицы	133
Миграция с непартиционированной таблицы на партиционированную.....	134
Исправление, восстановление и замена данных таблицы.....	135
Восстановление и замена таблиц.....	136
Удаление данных и очистка разделов.....	137
Жизненный цикл таблицы Delta Lake	138
Восстановление таблицы	138
Очистка.....	139
Заключение	141
Глава 6. Создание нативных приложений на базе Delta Lake	142
Начало работы.....	143
Python	143
Rust.....	154
Создание функций Lambda.....	158
Что дальше?.....	165
Глава 7. Потокковая загрузка и выгрузка данных в Delta Lake	166
Потокковая передача данных и Delta Lake	166
Потокковая обработка данных в сравнении с пакетной	167
Delta Lake как источник.....	173
Delta Lake как приемник	174
Конфигурационные опции потокковой передачи Delta	176
Ограничение скорости ввода	176
Игнорирование обновлений или удалений	178
Начальная позиция обработки.....	180
Исходный снимок состояния с использованием опции withEventTimeOrder	182
Расширенное использование Delta Lake с Apache Spark	184
Идемпотентные потоковые записи	184
Метрики производительности Delta Lake	189
Инструменты Auto Loader и Delta Live Tables.....	190
Оптимизированный источник Auto Loader (Автоматическая загрузка данных)	191
Декларативная среда Delta Live Tables (динамические таблицы Delta).....	191
Поток изменений данных (Change Data Feed).....	193
Использование Change Data Feed.....	194
Схема.....	198
Заключение	200

Глава 8. Расширенные возможности	201
Генерируемые столбцы, ключи и идентификаторы.....	201
Комментарии и ограничения.....	204
Комментарии.....	204
Ограничения для таблиц Delta	206
Векторы удаления	208
Стратегия Merge-on-Read.....	208
Пошаговая демонстрация работы векторов удаления	210
Заключение	214
Глава 9. Проектирование озера-хранилища данных.....	215
Архитектура Lakehouse (озеро-хранилище данных)	215
Что такое озеро-хранилище данных?	216
Из опыта традиционных хранилищ данных	216
Из опыта традиционных озер данных.....	217
Двухуровневая архитектура данных	219
Архитектура озера-хранилища данных.....	220
Основы работы с Delta Lake	222
Открытый исходный код на базе открытых стандартов в открытой экосистеме.....	222
Поддержка транзакций.....	224
Принудительное применение схемы и управление схемами данных	226
Архитектура медальона (Medallion)	230
Изучение уровня Bronze	231
Изучение уровня Silver	234
Изучение уровня Gold	238
Архитектура медальона для потоковой передачи данных.....	239
Заключение	241
Глава 10. Тонкая настройка производительности: оптимизация конвейеров данных с помощью Delta Lake.....	243
Цели повышения производительности	244
Максимизация производительности чтения	244
Максимизация производительности записи	246
Аспекты производительности.....	248
Партиционирование	248
Служебные функции таблиц	250
Статистика таблиц.....	258

Параметр CLUSTER BY	267
Индекс на основе фильтров Блума	272
Заключение	274
Глава 11. Успешные паттерны проектирования	275
Радикальное сокращение ресурсозатрат на вычисления	275
Высокоскоростные решения	276
Интеграция умных устройств	276
Эффективная потоковая загрузка данных	284
Потоковая загрузка данных	284
Зарождение Delta Rust	286
Эволюция загрузки данных	287
Координация сложных систем	290
Объединение оперативных хранилищ данных в DoorDash	290
Захват изменений данных	291
Совместная работа Delta и Flink	293
Заключение	295
Глава 12. Основы управления и безопасности озера-хранилища данных	296
Управление данными в озере-хранилище	297
Зарождение принципов управления данными	304
Продукты данных и их взаимосвязь с активами данных	306
Продукты данных в архитектуре Lakehouse	308
Поддержание высокого уровня доверия	308
Активы данных и доступ к ним	309
Модель активов данных	309
Унификация управления для озер и хранилищ данных	312
Управление разрешениями	313
Разрешения файловой системы	314
Управление доступом к облачному объектному хранилищу	316
Управление идентификацией и доступом	316
Безопасность данных	318
Детализированное управление доступом для озера-хранилища	330
Заключение	332
Глава 13. Управление метаданными, потоки данных и происхождение данных	333
Управление метаданными	333
Что такое управление метаданными?	334
Каталоги данных	334

Надежность данных, распорядители данных и управление разрешениями.....	335
Почему так важно хранилище метаданных (Metastore)	336
Unity Catalog.....	338
Поток данных и их происхождение	341
Отслеживание происхождения	342
Обмен данными	348
Автоматизация жизненных циклов данных.....	348
Ведение журналов аудита.....	351
Мониторинг и оповещение.....	353
Что такое поиск и обнаружение данных?	354
Заключение	356
Глава 14. Обмен данными с помощью протокола Delta Sharing	357
Основы Delta Sharing	358
Поставщики данных.....	359
Получатели данных	360
Сервер Delta Sharing	362
REST API.....	362
Структура REST URI	363
Получение списка общих ресурсов	364
Получение сведений об общем ресурсе.....	365
Получение списка схем в общем ресурсе	366
Получение списка всех таблиц в общем ресурсе	370
Клиенты Delta Sharing.....	371
Delta Sharing с Apache Spark	371
Потоковая обработка данных с использованием Delta Sharing	375
Коннекторы от сообщества Delta Sharing.....	378
Заключение	378
Об авторах	380
Иллюстрация на обложке.....	382