

Научная библиотека

БНТУ



# Google BigQuery

## Всё о хранилищах данных, аналитике и машинном обучении

Валиappa Лакшманан

Джордан Тайджани

**НАВУКОВАЯ БІБЛІЯТЭКА**

Беларускага нацыянальнага  
тэхнічнага ўніверсітэта

Інв. №

**1883828**



Санкт-Петербург • Москва • Екатеринбург • Воронеж  
Нижний Новгород • Ростов-на-Дону  
Самара • Минск

2021

# Оглавление

<b>Отзывы о книге «Google BigQuery. Всё о хранилищах данных, аналитике и машинном обучении»</b> .....	<b>13</b>
<b>Предисловие</b> .....	<b>15</b>
Для кого написана эта книга? .....	16
Условные обозначения .....	16
Использование примеров программного кода .....	17
Благодарности .....	17
От издательства .....	18
<b>Глава 1. Что такое Google BigQuery?</b> .....	<b>19</b>
Архитектуры обработки данных .....	19
Система управления реляционными базами данных .....	20
Фреймворк MapReduce .....	22
BigQuery: бессерверный распределенный движок SQL .....	23
Работа с BigQuery .....	25
Анализ наборов данных .....	25
ETL, EL и ELT .....	26
Эффективная аналитика .....	28
Простота управления .....	30
История появления BigQuery .....	31
Что позволило создать BigQuery? .....	34
Отделение вычислений от хранилища .....	35
Хранилище и сетевая инфраструктура .....	36
Управляемое хранилище .....	37
Интеграция с платформой Google Cloud .....	39
Безопасность и соответствие требованиям .....	40
Выводы .....	41

<b>Глава 2. Основы запросов .....</b>	<b>42</b>
Простые запросы.....	44
Извлечение записей с помощью SELECT.....	44
Создание псевдонимов столбцов с помощью AS .....	46
Фильтрация с WHERE.....	48
SELECT *, EXCEPT, REPLACE.....	49
Подзапросы с WITH .....	50
Сортировка с ORDER BY.....	50
Агрегирование.....	51
Агрегирование с GROUP BY.....	51
Подсчет записей с COUNT .....	52
Фильтрация сгруппированных значений с HAVING.....	53
Поиск уникальных значений с DISTINCT.....	54
Краткое руководство по массивам и структурам.....	55
Создание массивов с помощью ARRAY_AGG.....	57
Массив структур STRUCT .....	59
Кортежи .....	60
Работа с массивами.....	60
Развертывание массива .....	61
Соединение таблиц .....	62
Основы соединения таблиц.....	63
Оператор внутреннего соединения INNER JOIN.....	66
Оператор перекрестного соединения CROSS JOIN.....	67
Оператор внешнего соединения OUTER JOIN.....	69
Сохранение и совместное использование.....	70
История запросов и кеширование .....	70
Сохранение запросов .....	72
Представления и общедоступные запросы.....	73
Выводы .....	74
<b>Глава 3. Типы данных, функции и операторы.....</b>	<b>75</b>
Числовые типы и функции.....	76
Математические функции.....	77
Стандартное вещественное деление.....	78

Функции SAFE .....	78
Сравнение .....	79
Точные десятичные вычисления с NUMERIC .....	80
Тип BOOL .....	81
Логические операции .....	81
Условные выражения .....	83
Обработка NULL с помощью COALESCE .....	84
Явное и неявное приведение типов .....	85
Использование COUNTIF, чтобы избежать приведения логических значений .....	87
Строковые функции .....	88
Интернационализация .....	89
Формирование и парсинг строк .....	91
Функции для обработки строк .....	91
Функции преобразования .....	92
Регулярные выражения .....	92
Краткие итоги по строковым функциям .....	94
Операции со значениями TIMESTAMP .....	95
Парсинг и форматирование отметок времени .....	95
Извлечение календарных данных .....	97
Арифметические операции с временными метками .....	98
DATE, TIME и DATETIME .....	99
Функции для работы с географическими координатами .....	100
Выводы .....	101
<b>Глава 4. Загрузка данных в BigQuery .....</b>	<b>104</b>
Основы .....	104
Загрузка из локального источника .....	105
Корректировка схемы .....	112
Копирование в новую таблицу .....	116
Управление данными (DDL и DML) .....	116
Эффективная загрузка данных .....	118
Федеративные запросы и внешние источники данных .....	121
Как использовать федеративные запросы .....	121
Когда использовать федеративные запросы и внешние источники данных .....	125

Интерактивное исследование и запрос данных из Google Sheets .....	132
Запросы SQL для выборки данных из Cloud Bigtable .....	141
Передача и экспорт данных.....	147
Служба передачи данных Data Transfer Service.....	147
Экспортирование журналов Stackdriver .....	153
Использование Cloud Dataflow для чтения/записи в BigQuery.....	154
Перемещение локальных данных.....	159
Методы миграции данных .....	159
Выводы .....	162
<b>Глава 5. Разработка с BigQuery .....</b>	<b>163</b>
Программный доступ .....	163
Доступ к BigQuery через REST API.....	163
Google Cloud Client Library.....	171
Доступ к BigQuery из инструментов исследования данных .....	188
Блокноты в Google Cloud Platform .....	188
Работа с BigQuery, pandas и Jupyter.....	193
Работа с BigQuery из R.....	198
Cloud Dataflow .....	199
Драйверы JDBC/ODBC.....	202
Внедрение данных из BigQuery в Google Slides (в G Suite) .....	203
Bash-скрипты для BigQuery.....	205
Создание наборов данных и таблиц .....	205
Выполнение запросов .....	208
Объекты BigQuery.....	210
Выводы .....	212
<b>Глава 6. Архитектура BigQuery .....</b>	<b>213</b>
Архитектура высокого уровня .....	213
Жизненный цикл запроса .....	213
Обновление BigQuery .....	218
Система обработки запросов (Dremel).....	219
Архитектура Dremel.....	221
Выполнение запроса .....	226
Хранилище.....	241
Хранение данных .....	241

Метаданные .....	248
Выводы .....	258
<b>Глава 7. Оптимизация производительности и затрат .....</b>	<b>259</b>
Принципы производительности .....	259
Ключевые составляющие производительности .....	260
Управление затратами .....	260
Измерение производительности и поиск проблем .....	262
Определение скорости выполнения запроса с помощью REST API .....	263
Определение скорости выполнения запроса с помощью BigQuery Workload Tester .....	265
Выявление проблем в рабочих нагрузках с помощью Stackdriver .....	267
Чтение плана запроса .....	269
Увеличение скорости выполнения запросов .....	274
Минимизация ввода/вывода .....	275
Кеширование результатов предыдущих запросов .....	280
Эффективное выполнение соединений .....	284
Исключение перегрузки рабочих серверов .....	293
Использование приближенных функций агрегирования .....	296
Оптимизация хранения данных и доступа к ним .....	299
Минимизация сетевых издержек .....	300
Выбор эффективного формата хранения .....	303
Секционирование таблиц для уменьшения объема сканирования .....	313
Кластеризация таблиц на основе ключей с большой мощностью множества .....	316
Случаи использования, нечувствительные ко времени .....	321
Пакетные запросы .....	321
Загрузка файлов .....	323
Выводы .....	324
Контрольный список .....	324
<b>Глава 8. Продвинутое использование запросов .....</b>	<b>326</b>
Многократные запросы .....	326
Параметризованные запросы .....	327
Пользовательские функции SQL .....	332
Повторное использование частей запросов .....	337

Продвинутый SQL .....	341
Работа с массивами .....	342
Оконные функции .....	351
Метаданные таблиц .....	356
Язык определения данных и язык манипулирования данными .....	360
За пределами SQL .....	365
Пользовательские функции на JavaScript .....	366
Скрипты .....	367
Продвинутые функции .....	375
Геоинформационная система BigQuery .....	375
Полезные статистические функции .....	383
Алгоритмы хеширования .....	385
Выводы .....	389
<b>Глава 9. Машинное обучение в BigQuery .....</b>	<b>390</b>
Что такое машинное обучение? .....	390
Формулировка задачи машинного обучения .....	391
Типы задач машинного обучения .....	392
Построение регрессионной модели .....	396
Выбор метки .....	396
Выбор признаков в наборе данных .....	397
Создание обучающего набора данных .....	401
Обучение и оценка модели .....	402
Получение прогнозов с помощью модели .....	404
Исследование весов модели .....	407
Более сложные регрессионные модели .....	409
Создание модели классификации .....	414
Обучение .....	415
Оценка .....	416
Прогнозирование .....	417
Выбор порога .....	418
Настройка механизма машинного обучения в BigQuery .....	420
Управление делением данных .....	420
Балансировка классов .....	422
Регуляризация .....	422

Классификация методом k-средних.....	423
Выбор признаков для классификации.....	424
Классификация пунктов проката велосипедов.....	425
Классификация.....	426
Исследование кластеров.....	427
Принятие решений на основе данных.....	429
Рекомендательные системы.....	430
Набор данных MovieLens.....	430
Разложение матрицы.....	432
Получение рекомендаций.....	434
Включение информации о пользователях и фильмах.....	436
Нестандартные модели машинного обучения в GCP.....	443
Настройка гиперпараметров.....	444
AutoML.....	448
Поддержка TensorFlow.....	450
Выводы.....	453
<b>Глава 10. Администрирование и безопасность BigQuery.....</b>	<b>455</b>
Защищенность инфраструктуры.....	455
Управление идентификацией и доступом.....	457
Идентификация.....	457
Роль.....	458
Ресурс.....	461
Администрирование BigQuery.....	462
Управление заданиями.....	462
Авторизация пользователей.....	463
Восстановление удаленных записей и таблиц.....	463
Непрерывная интеграция/непрерывное развертывание.....	464
Экспорт биллинга — получение информации о расходах.....	467
Оперативные панели, мониторинг и журналы аудита.....	470
Доступность, восстановление после отказа и шифрование.....	471
Зоны, регионы и объединения регионов.....	471
BigQuery и обработка отказов.....	472
Сохранность, резервное копирование и восстановление после аварий.....	476
Конфиденциальность и шифрование.....	477



Соответствие требованиям законодательств .....	478
Местоположение данных .....	478
Ограничение доступа к подмножествам данных .....	480
Удаление всех сделок, связанных с конкретным физическим лицом .....	483
Предотвращение потери данных .....	487
СМЕК .....	488
Защита от утечки данных .....	490
Выводы .....	491
<b>Об авторах .....</b>	<b>493</b>
<b>Об обложке .....</b>	<b>494</b>