

Google BigQuery The Definitive Guide

1. Preface
 - a. Who Is This Book For?
 - b. Conventions Used in This Book
 - c. Using Code Examples
 - d. O'Reilly Online Learning
 - e. How to Contact Us
 - f. Acknowledgments
2. 1. What Is Google BigQuery?
 - a. Data Processing Architectures
 - i. Relational Database Management System
 - ii. MapReduce Framework
 - iii. BigQuery: A Serverless, Distributed SQL Engine
 - b. Working with BigQuery
 - i. Deriving Insights Across Datasets
 - ii. ETL, EL, and ELT
 - iii. Powerful Analytics
 - iv. Simplicity of Management
 - c. How BigQuery Came About
 - d. What Makes BigQuery Possible?
 - i. Separation of Compute and Storage
 - ii. Storage and Networking Infrastructure
 - iii. Managed Storage
 - iv. Integration with Google Cloud Platform
 - v. Security and Compliance
 - e. Summary
3. 2. Query Essentials
 - a. Simple Queries
 - i. Retrieving Rows by Using SELECT
 - ii. Aliasing Column Names with AS
 - iii. Filtering with WHERE
 - iv. SELECT *, EXCEPT, REPLACE
 - v. Subqueries with WITH
 - vi. Sorting with ORDER BY
 - b. Aggregates
 - i. Computing Aggregates by Using GROUP BY
 - ii. Counting Records by Using COUNT
 - iii. Filtering Grouped Items by Using HAVING
 - iv. Finding Unique Values by Using DISTINCT
 - c. A Brief Primer on Arrays and Structs
 - i. Creating Arrays by Using ARRAY_AGG
 - ii. Array of STRUCT
 - iii. TUPLE
 - iv. Working with Arrays
 - v. UNNEST an Array
 - d. Joining Tables
 - i. The JOIN Explained

- ii. INNER JOIN
 - iii. CROSS JOIN
 - iv. OUTER JOIN
- e. Saving and Sharing
 - i. Query History and Caching
 - ii. Saved Queries
 - iii. Views Versus Shared Queries
- f. Summary
- 4. 3. Data Types, Functions, and Operators
 - a. Numeric Types and Functions
 - i. Mathematical Functions
 - ii. Standard-Compliant Floating-Point Division
 - iii. SAFE Functions
 - iv. Comparisons
 - v. Precise Decimal Calculations with NUMERIC
 - b. Working with BOOL
 - i. Logical Operations
 - ii. Conditional Expressions
 - iii. Cleaner NULL-Handling with COALESCE
 - iv. Casting and Coercion
 - v. Using COUNTIF to Avoid Casting Booleans
 - c. String Functions
 - i. Internationalization
 - ii. Printing and Parsing
 - iii. String Manipulation Functions
 - iv. Transformation Functions
 - v. Regular Expressions
 - vi. Summary of String Functions
 - d. Working with TIMESTAMP
 - i. Parsing and Formatting Timestamps
 - ii. Extracting Calendar Parts
 - iii. Arithmetic with Timestamps
 - iv. Date, Time, and DateTime
 - e. Working with GIS Functions
 - f. Summary
- 5. 4. Loading Data into BigQuery
 - a. The Basics
 - i. Loading from a Local Source
 - ii. Specifying a Schema
 - iii. Copying into a New Table
 - iv. Data Management (DDL and DML)
 - v. Loading Data Efficiently
 - b. Federated Queries and External Data Sources
 - i. How to Use Federated Queries
 - ii. When to Use Federated Queries and External Data Sources
 - iii. Interactive Exploration and Querying of Data in Google Sheets
 - iv. SQL Queries on Data in Cloud Bigtable
 - c. Transfers and Exports
 - i. Data Transfer Service

- ii. Exporting Stackdriver Logs
 - iii. Using Cloud Dataflow to Read/Write from BigQuery
 - d. Moving On-Premises Data
 - i. Data Migration Methods
 - e. Summary
- 6. 5. Developing with BigQuery
 - a. Developing Programmatically
 - i. Accessing BigQuery via the REST API
 - ii. Google Cloud Client Library
 - b. Accessing BigQuery from Data Science Tools
 - i. Notebooks on Google Cloud Platform
 - ii. Working with BigQuery, pandas, and Jupyter
 - iii. Working with BigQuery from R
 - iv. Cloud Dataflow
 - v. JDBC/ODBC drivers
 - vi. Incorporating BigQuery Data into Google Slides (in G Suite)
 - c. Bash Scripting with BigQuery
 - i. Creating Datasets and Tables
 - ii. Executing Queries
 - iii. BigQuery Objects
 - d. Summary
- 7. 6. Architecture of BigQuery
 - a. High-Level Architecture
 - i. Life of a Query Request
 - ii. BigQuery Upgrades
 - b. Query Engine (Dremel)
 - i. Dremel Architecture
 - ii. Query Execution
 - c. Storage
 - i. Storage Data
 - ii. Metadata
 - d. Summary
- 8. 7. Optimizing Performance and Cost
 - a. Principles of Performance
 - i. Key Drivers of Performance
 - ii. Controlling Cost
 - b. Measuring and Troubleshooting
 - i. Measuring Query Speed Using REST API
 - ii. Measuring Query Speed Using BigQuery Workload Tester
 - iii. Troubleshooting Workloads Using Stackdriver
 - iv. Reading Query Plan Information
 - c. Increasing Query Speed
 - i. Minimizing I/O
 - ii. Caching the Results of Previous Queries
 - iii. Performing Efficient Joins
 - iv. Avoiding Overwhelming a Worker
 - v. Using Approximate Aggregation Functions
 - d. Optimizing How Data Is Stored and Accessed
 - i. Minimizing Network Overhead

- ii. Choosing an Efficient Storage Format
 - iii. Partitioning Tables to Reduce Scan Size
 - iv. Clustering Tables Based on High-Cardinality Keys
 - e. Time-Insensitive Use Cases
 - i. Batch Queries
 - ii. File Loads
 - f. Summary
 - i. Checklist
- 9. 8. Advanced Queries
 - a. Reusable Queries
 - i. Parameterized Queries
 - ii. SQL User-Defined Functions
 - iii. Reusing Parts of Queries
 - b. Advanced SQL
 - i. Working with Arrays
 - ii. Window Functions
 - iii. Table Metadata
 - iv. Data Definition Language and Data Manipulation Language
 - c. Beyond SQL
 - i. JavaScript UDFs
 - ii. Scripting
 - d. Advanced Functions
 - i. BigQuery Geographic Information Systems
 - ii. Useful Statistical Functions
 - iii. Hash Algorithms
 - e. Summary
- 10. 9. Machine Learning in BigQuery
 - a. What Is Machine Learning?
 - i. Formulating a Machine Learning Problem
 - ii. Types of Machine Learning Problems
 - b. Building a Regression Model
 - i. Choose the Label
 - ii. Exploring the Dataset to Find Features
 - iii. Creating a Training Dataset
 - iv. Training and Evaluating the Model
 - v. Predicting with the Model
 - vi. Examining Model Weights
 - vii. More-Complex Regression Models
 - c. Building a Classification Model
 - i. Training
 - ii. Evaluation
 - iii. Prediction
 - iv. Choosing the Threshold
 - d. Customizing BigQuery ML
 - i. Controlling Data Split
 - ii. Balancing Classes
 - iii. Regularization
 - e. k-Means Clustering
 - i. What's Being Clustered?

- ii. Clustering Bicycle Stations
 - iii. Carrying Out Clustering
 - iv. Understanding the Clusters
 - v. Data-Driven Decisions
- f. Recommender Systems
 - i. The MovieLens Dataset
 - ii. Matrix Factorization
 - iii. Making Recommendations
 - iv. Incorporating User and Movie Information
- g. Custom Machine Learning Models on GCP
 - i. Hyperparameter Tuning
 - ii. AutoML
 - iii. Support for TensorFlow
- h. Summary
- 11. 10. Administering and Securing BigQuery
 - a. Infrastructure Security
 - b. Identity and Access Management
 - i. Identity
 - ii. Role
 - iii. Resource
 - c. Administering BigQuery
 - i. Job Management
 - ii. Authorizing Users
 - iii. Restoring Deleted Records and Tables
 - iv. Continuous Integration/Continuous Deployment
 - v. Cost/Billing Exports
 - vi. Dashboards, Monitoring, and Audit Logging
 - d. Availability, Disaster Recovery, and Encryption
 - i. Zones, Regions, and Multiregions
 - ii. BigQuery and Failure Handling
 - iii. Durability, Backups, and Disaster Recovery
 - iv. Privacy and Encryption
 - e. Regulatory Compliance
 - i. Data Locality
 - ii. Restricting Access to Subsets of Data
 - iii. Removing All Transactions Related to a Single Individual
 - iv. Data Loss Prevention
 - v. CMEK
 - vi. Data Exfiltration Protection
 - f. Summary
- 12. Index