

# Table of Contents

|                                                |             |
|------------------------------------------------|-------------|
| <b>Preface</b> .....                           | <b>xiii</b> |
| <b>1. What Is Google BigQuery?</b> .....       | <b>1</b>    |
| Data Processing Architectures                  | 1           |
| Relational Database Management System          | 2           |
| MapReduce Framework                            | 3           |
| BigQuery: A Serverless, Distributed SQL Engine | 5           |
| Working with BigQuery                          | 7           |
| Deriving Insights Across Datasets              | 7           |
| ETL, EL, and ELT                               | 8           |
| Powerful Analytics                             | 10          |
| Simplicity of Management                       | 12          |
| How BigQuery Came About                        | 12          |
| What Makes BigQuery Possible?                  | 15          |
| Separation of Compute and Storage              | 15          |
| Storage and Networking Infrastructure          | 16          |
| Managed Storage                                | 18          |
| Integration with Google Cloud Platform         | 19          |
| Security and Compliance                        | 21          |
| Summary                                        | 21          |
| <b>2. Query Essentials</b> .....               | <b>23</b>   |
| Simple Queries                                 | 24          |
| Retrieving Rows by Using SELECT                | 25          |
| Aliasing Column Names with AS                  | 27          |
| Filtering with WHERE                           | 28          |
| SELECT *, EXCEPT, REPLACE                      | 29          |

|                                                |           |
|------------------------------------------------|-----------|
| Subqueries with WITH                           | 30        |
| Sorting with ORDER BY                          | 31        |
| Aggregates                                     | 31        |
| Computing Aggregates by Using GROUP BY         | 31        |
| Counting Records by Using COUNT                | 32        |
| Filtering Grouped Items by Using HAVING        | 33        |
| Finding Unique Values by Using DISTINCT        | 33        |
| A Brief Primer on Arrays and Structs           | 35        |
| Creating Arrays by Using ARRAY_AGG             | 36        |
| Array of STRUCT                                | 39        |
| TUPLE                                          | 39        |
| Working with Arrays                            | 40        |
| UNNEST an Array                                | 41        |
| Joining Tables                                 | 42        |
| The JOIN Explained                             | 42        |
| INNER JOIN                                     | 45        |
| CROSS JOIN                                     | 46        |
| OUTER JOIN                                     | 47        |
| Saving and Sharing                             | 48        |
| Query History and Caching                      | 49        |
| Saved Queries                                  | 50        |
| Views Versus Shared Queries                    | 51        |
| Summary                                        | 51        |
| <b>3. Data Types, Functions, and Operators</b> | <b>53</b> |
| Numeric Types and Functions                    | 54        |
| Mathematical Functions                         | 55        |
| Standard-Compliant Floating-Point Division     | 55        |
| SAFE Functions                                 | 56        |
| Comparisons                                    | 56        |
| Precise Decimal Calculations with NUMERIC      | 57        |
| Working with BOOL                              | 59        |
| Logical Operations                             | 59        |
| Conditional Expressions                        | 60        |
| Cleaner NULL-Handling with COALESCE            | 61        |
| Casting and Coercion                           | 62        |
| Using COUNTIF to Avoid Casting Booleans        | 64        |
| String Functions                               | 65        |
| Internationalization                           | 66        |
| Printing and Parsing                           | 67        |
| String Manipulation Functions                  | 68        |

|                                                               |            |
|---------------------------------------------------------------|------------|
| Transformation Functions                                      | 68         |
| Regular Expressions                                           | 69         |
| Summary of String Functions                                   | 70         |
| Working with TIMESTAMP                                        | 71         |
| Parsing and Formatting Timestamps                             | 71         |
| Extracting Calendar Parts                                     | 73         |
| Arithmetic with Timestamps                                    | 74         |
| Date, Time, and DateTime                                      | 74         |
| Working with GIS Functions                                    | 75         |
| Summary                                                       | 76         |
| <b>4. Loading Data into BigQuery</b>                          | <b>79</b>  |
| The Basics                                                    | 79         |
| Loading from a Local Source                                   | 80         |
| Specifying a Schema                                           | 87         |
| Copying into a New Table                                      | 90         |
| Data Management (DDL and DML)                                 | 90         |
| Loading Data Efficiently                                      | 92         |
| Federated Queries and External Data Sources                   | 95         |
| How to Use Federated Queries                                  | 95         |
| When to Use Federated Queries and External Data Sources       | 99         |
| Interactive Exploration and Querying of Data in Google Sheets | 105        |
| SQL Queries on Data in Cloud Bigtable                         | 114        |
| Transfers and Exports                                         | 119        |
| Data Transfer Service                                         | 119        |
| Exporting Cloud Logging Logs                                  | 125        |
| Using Cloud Dataflow to Read/Write from BigQuery              | 127        |
| Moving On-Premises Data                                       | 131        |
| Data Migration Methods                                        | 132        |
| Summary                                                       | 134        |
| <b>5. Developing with BigQuery</b>                            | <b>137</b> |
| Developing Programmatically                                   | 137        |
| Accessing BigQuery via the REST API                           | 137        |
| Google Cloud Client Library                                   | 144        |
| Accessing BigQuery from Data Science Tools                    | 161        |
| Notebooks on Google Cloud Platform                            | 161        |
| Working with BigQuery, pandas, and Jupyter                    | 165        |
| Working with BigQuery from R                                  | 171        |
| Cloud Dataflow                                                | 172        |
| JDBC/ODBC drivers                                             | 174        |

|                                                             |            |
|-------------------------------------------------------------|------------|
| Incorporating BigQuery Data into Google Slides (in G Suite) | 175        |
| Bash Scripting with BigQuery                                | 177        |
| Creating Datasets and Tables                                | 178        |
| Executing Queries                                           | 181        |
| BigQuery Objects                                            | 183        |
| Summary                                                     | 185        |
| <b>6. Architecture of BigQuery</b>                          | <b>187</b> |
| High-Level Architecture                                     | 187        |
| Life of a Query Request                                     | 187        |
| BigQuery Upgrades                                           | 192        |
| Query Engine (Dremel)                                       | 192        |
| Dremel Architecture                                         | 194        |
| Query Execution                                             | 199        |
| Storage                                                     | 214        |
| Storage Data                                                | 214        |
| Metadata                                                    | 220        |
| Summary                                                     | 229        |
| <b>7. Optimizing Performance and Cost</b>                   | <b>231</b> |
| Principles of Performance                                   | 231        |
| Key Drivers of Performance                                  | 232        |
| Controlling Cost                                            | 232        |
| Measuring and Troubleshooting                               | 234        |
| Measuring Query Speed Using REST API                        | 235        |
| Measuring Query Speed Using BigQuery Workload Tester        | 236        |
| Troubleshooting Workloads Using Cloud Logging               | 238        |
| Reading Query Plan Information                              | 240        |
| Increasing Query Speed                                      | 246        |
| Minimizing I/O                                              | 247        |
| Caching the Results of Previous Queries                     | 253        |
| Performing Efficient Joins                                  | 257        |
| Avoiding Overwhelming a Worker                              | 266        |
| Using Approximate Aggregation Functions                     | 269        |
| Optimizing How Data Is Stored and Accessed                  | 272        |
| Minimizing Network Overhead                                 | 272        |
| Choosing an Efficient Storage Format                        | 276        |
| Partitioning Tables to Reduce Scan Size                     | 285        |
| Clustering Tables Based on High-Cardinality Keys            | 289        |
| Cost Optimization                                           | 293        |
| Batch Queries                                               | 293        |

|                                                         |            |
|---------------------------------------------------------|------------|
| Optimizing Storage Cost                                 | 294        |
| Query Cost Control                                      | 295        |
| File Loads                                              | 296        |
| Summary                                                 | 297        |
| Checklist                                               | 297        |
| <b>8. Advanced Queries</b>                              | <b>299</b> |
| Reusable Queries                                        | 299        |
| Parameterized Queries                                   | 300        |
| SQL User-Defined Functions                              | 305        |
| Reusing Parts of Queries                                | 309        |
| Advanced SQL                                            | 313        |
| Working with Arrays                                     | 314        |
| Window Functions                                        | 322        |
| Table Metadata                                          | 328        |
| Data Definition Language and Data Manipulation Language | 331        |
| Beyond SQL                                              | 336        |
| JavaScript UDFs                                         | 336        |
| Scripting                                               | 338        |
| Advanced Functions                                      | 349        |
| BigQuery Geographic Information Systems                 | 349        |
| Useful Statistical Functions                            | 356        |
| Hash Algorithms                                         | 359        |
| Summary                                                 | 362        |
| <b>9. Machine Learning in BigQuery</b>                  | <b>365</b> |
| What Is Machine Learning?                               | 365        |
| Formulating a Machine Learning Problem                  | 366        |
| Types of Machine Learning Problems                      | 368        |
| Building a Regression Model                             | 371        |
| Choose the Label                                        | 371        |
| Exploring the Dataset to Find Features                  | 372        |
| Creating a Training Dataset                             | 376        |
| Training and Evaluating the Model                       | 377        |
| Predicting with the Model                               | 379        |
| Examining Model Weights                                 | 381        |
| More-Complex Regression Models                          | 383        |
| Building a Classification Model                         | 393        |
| Training                                                | 393        |
| Evaluation                                              | 395        |
| Prediction                                              | 396        |

|                                                     |            |
|-----------------------------------------------------|------------|
| Choosing the Threshold                              | 397        |
| Customizing BigQuery ML                             | 399        |
| Controlling Data Split                              | 399        |
| Balancing Classes                                   | 400        |
| Regularization                                      | 401        |
| k-Means Clustering                                  | 402        |
| What's Being Clustered?                             | 402        |
| Clustering Bicycle Stations                         | 403        |
| Carrying Out Clustering                             | 405        |
| Understanding the Clusters                          | 405        |
| Data-Driven Decisions                               | 407        |
| Recommender Systems                                 | 408        |
| The MovieLens Dataset                               | 408        |
| Matrix Factorization                                | 410        |
| Making Recommendations                              | 412        |
| Incorporating User and Movie Information            | 414        |
| Custom Machine Learning Models on GCP               | 421        |
| Hyperparameter Tuning                               | 421        |
| AutoML                                              | 426        |
| Support for TensorFlow                              | 427        |
| Summary                                             | 432        |
| <b>10. Administering and Securing BigQuery.....</b> | <b>435</b> |
| Infrastructure Security                             | 435        |
| Identity and Access Management                      | 437        |
| Identity                                            | 437        |
| Role                                                | 438        |
| Resource                                            | 441        |
| Administering BigQuery                              | 441        |
| Job Management                                      | 441        |
| Authorizing Users                                   | 442        |
| Restoring Deleted Records and Tables                | 442        |
| Continuous Integration/Continuous Deployment        | 443        |
| Slot Reservations                                   | 446        |
| Cost/Billing Exports                                | 448        |
| Dashboards, Monitoring, and Audit Logging           | 450        |
| Availability, Disaster Recovery, and Encryption     | 452        |
| Zones, Regions, and Multiregions                    | 452        |
| BigQuery and Failure Handling                       | 453        |
| Durability, Backups, and Disaster Recovery          | 456        |
| Privacy and Encryption                              | 456        |

|                                                          |     |
|----------------------------------------------------------|-----|
| Regulatory Compliance                                    | 458 |
| Data Locality                                            | 458 |
| Restricting Access to Subsets of Data                    | 459 |
| Removing All Transactions Related to a Single Individual | 463 |
| Data Loss Prevention                                     | 467 |
| CMEK                                                     | 469 |
| Data Exfiltration Protection                             | 470 |
| Summary                                                  | 471 |

|                   |            |
|-------------------|------------|
| <b>Index.....</b> | <b>473</b> |
|-------------------|------------|

Enterprises are becoming increasingly data driven, and a key component of any enterprise's data strategy is a data warehouse—a central repository of integrated data from all across the company. Traditionally, the data warehouse was used by data analysts to create analytical reports. But now it is also increasingly used to populate real-time dashboards, to make ad hoc queries, and to provide decision-making guidance through predictive analytics. Because of these business requirements for advanced analytics and a trend toward cost control, agility, and self-service data access, many organizations are moving to cloud-based data warehouses such as Google BigQuery.

In this book, we provide a thorough tour of BigQuery, a serverless, highly scalable, low-cost enterprise data warehouse that is available on Google Cloud. Because there is no infrastructure to manage, enterprises can focus on analyzing data to find meaningful insights using familiar SQL.

Our goal with BigQuery has been to build a data platform that provides leading-edge capabilities, takes advantage of the many great technologies that are now available in cloud environments, and supports tried-and-true data technologies that are still relevant today. For example, on the leading edge, Google's BigQuery is a serverless compute architecture that decouples compute and storage. This enables diverse layers of the architecture to perform and scale independently, and it gives data developers flexibility in design and deployment. Somewhat uniquely, BigQuery supports native machine learning and geospatial analysis. With Cloud Pub/Sub, Cloud Dataflow, Cloud Bigtable, Cloud AI Platform, and many third-party integrations, BigQuery interoperates with both traditional and modern systems, at a wide range of desired throughput and latency. And on the tried-and-true front, BigQuery supports ANSI-standard SQL, columnar optimization, and federated queries, which are key to the self-service ad hoc data exploration that many users demand.