

---

# Table of Contents

|   |            |
|---|------------|
| <b>Introduction.....</b>                          | <b>vii</b> |
| <b>1. Why Impala?.....</b>                        | <b>1</b>   |
| Impala's Place in the Big Data Ecosystem          | 1          |
| Flexibility for Your Big Data Workflow            | 2          |
| High-Performance Analytics                        | 3          |
| Exploratory Business Intelligence                 | 3          |
| <b>2. Getting Up and Running with Impala.....</b> | <b>5</b>   |
| Installation                                      | 5          |
| Connecting to Impala                              | 6          |
| Your First Impala Queries                         | 7          |
| <b>3. Impala for the Database Developer.....</b>  | <b>11</b>  |
| The SQL Language                                  | 12         |
| Standard SQL                                      | 12         |
| Limited DML                                       | 12         |
| No Transactions                                   | 13         |
| Numbers   | 13         |
| Recent Additions                                  | 14         |
| Big Data Considerations                           | 15         |
| Billions and Billions of Rows                     | 15         |
| HDFS Block Size                                   | 16         |
| Parquet Files: The Biggest Blocks of All          | 16         |
| How Impala Is Like a Data Warehouse               | 17         |
| Physical and Logical Data Layouts                 | 18         |
| The HDFS Storage Model                            | 18         |
| Distributed Queries                               | 19         |
| Normalized and Denormalized Data                  | 21         |

|   |           |
|---|-----------|
| File Formats  | 21        |
| Text File Format                                      | 22        |
| Parquet File Format                                   | 23        |
| Getting File Format Information                       | 25        |
| Switching File Formats                                | 25        |
| Aggregation   | 26        |
| <b>4. Common Developer Tasks for Impala</b> .....     | <b>27</b> |
| Getting Data into an Impala Table                     | 27        |
| INSERT Statement                                      | 28        |
| LOAD DATA Statement                                   | 28        |
| External Tables                                       | 29        |
| Figuring Out Where Impala Data Resides                | 29        |
| Manually Loading Data Files into HDFS                 | 30        |
| Hive  | 30        |
| Sqoop   | 31        |
| Kite  | 31        |
| Porting SQL Code to Impala                            | 32        |
| Using Impala from a JDBC or ODBC Application          | 32        |
| JDBC  | 33        |
| ODBC  | 33        |
| Using Impala with a Scripting Language                | 34        |
| Running Impala SQL Statements from Scripts            | 34        |
| Variable Substitution                                 | 34        |
| Saving Query Results                                  | 35        |
| The impyla Package for Python Scripting               | 35        |
| Optimizing Impala Performance                         | 36        |
| Optimizing Query Performance                          | 37        |
| Optimizing Memory Usage                               | 37        |
| Working with Partitioned Tables                       | 39        |
| Finding the Ideal Granularity                         | 40        |
| Inserting into Partitioned Tables                     | 40        |
| Adding and Loading New Partitions                     | 41        |
| Writing User-Defined Functions                        | 42        |
| Collaborating with Your Administrators                | 43        |
| Designing for Security                                | 43        |
| Understanding Resource Management                     | 44        |
| Helping to Plan for Performance (Stats, HDFS Caching) | 44        |
| Understanding Cluster Topology                        | 45        |
| Always Close Your Queries                             | 45        |

|   |           |
|---|-----------|
| <b>5. Tutorials and Deep Dives.....</b>         | <b>47</b> |
| Tutorial: From Unix Data File to Impala Table   | 47        |
| Tutorial: Queries Without a Table               | 49        |
| Tutorial: The Journey of a Billion Rows         | 51        |
| Generating a Billion Rows of CSV Data           | 51        |
| Normalizing the Original Data                   | 57        |
| Converting to Parquet Format                    | 61        |
| Making a Partitioned Table                      | 64        |
| Next Steps                                      | 69        |
| Deep Dive: Joins and the Role of Statistics     | 69        |
| Creating a Million-Row Table to Join With       | 69        |
| Loading Data and Computing Stats                | 70        |
| Reviewing the EXPLAIN Plan                      | 71        |
| Trying a Real Query                             | 74        |
| The Story So Far                                | 78        |
| Final Join Query with 1B x 1M Rows              | 79        |
| Anti-Pattern: A Million Little Pieces           | 79        |
| Tutorial: Across the Fourth Dimension           | 81        |
| TIMESTAMP Data Type                             | 81        |
| Format Strings for Dates and Times              | 81        |
| Working with Individual Date and Time Fields    | 82        |
| Date and Time Arithmetic                        | 83        |
| Let's Solve the Y2K Problem                     | 84        |
| More Fun with Dates                             | 87        |
| Tutorial: Verbose and Quiet impala-shell Output | 88        |
| Tutorial: When Schemas Evolve                   | 89        |
| Numbers Versus Strings                          | 91        |
| Dealing with Out-of-Range Integers              | 92        |
| Tutorial: Levels of Abstraction                 | 95        |
| String Formatting                               | 95        |
| Temperature Conversion                          | 96        |