

Querying Databricks with Spark SQL

*Leverage SQL to query and
analyze Big Data for insights*

Adam Aspin



www.bpbonline.com

Copyright © 2024 BPB Online

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the author, nor BPB Online or its dealers and distributors, will be held liable for any damages caused or alleged to have been caused directly or indirectly by this book.

BPB Online has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, BPB Online cannot guarantee the accuracy of this information.

First published: 2024

Published by BPB Online

WeWork

119 Marylebone Road

London NW1 5PU

UK | UAE | INDIA | SINGAPORE

ISBN 978-93-55518-019

www.bpbonline.com

About the Author

Adam Aspin is an independent business intelligence consultant based in the United Kingdom. He has worked in Business Intelligence and analytics for over 25 years, and now focuses on Power BI. During this time, he has developed several dozen BI and analytics systems based on several different data platforms. Adam has been working with Databricks since it was first introduced, and has helped to deliver several analytical projects for multiple clients across Europe based on this technology.

Adam is a graduate of Oxford University. He has applied his skills for a range of clients in finance, banking, utilities, telecoms, construction, and retail. He is the author of a number of books, Querying MySQL, Querying MariaBD, Querying SQL Server, Querying Databricks with Spark SQL among others.

About the Reviewer

Max Mogenis is an accomplished professional in the data world. He has been focused on data architecture, data systems and professional consulting services for 30 plus years. Currently, he is serving as a Databricks Program lead and Enterprise Data Architect. Outside of work, he is passionate about motorcycles and traveling the globe.

Acknowledgement

My deepest gratitude goes yet again to the BPB production team for managing this, our sixth book, through the rocks and shoals of the publication process.

When delving into the arcane depths of technical products, it takes much work to maintain sight of the main objectives of a book. Fortunately, Max Mogenis, the technical reviewer, has worked unstintingly to help me retain focus on the purposes of this book. He has also shared his considerable experience of Databricks in the enterprise and has helped me immensely with his comments and suggestions.

Finally, my deepest gratitude has to be reserved for the two people who have given the most to this book. They are my wife and son, who have always encouraged me to persevere, providing all the support and encouragement that anyone could want. I am extremely lucky to have both of them.

Preface

Data is the lifeblood of the enterprise. Whether “Big Data” or the data stored in classic databases, it is queried using a variant of **Structured Query Language (SQL)**. So, simply put, SQL is key to data analysis. A mastery of SQL will help you to delve deep into the data that is stored in corporate databases. You can apply SQL to analyse the data and then present it in a clearly understandable form.

Databricks lets you use its flavour of SQL to query the data that this platform contains. This means that SQL can usually serve a vital role in preparing the data for final delivery, irrespective of the output application that you are using to present your analysis. Most end-user tools have an option for entering SQL to help derive meaning from the underlying data sources. Consequently, a knowledge of SQL can help you analyse data faster and more clearly. The aim of this book is to give you the necessary mastery of SQL, and specifically the Databricks flavour of SQL (known as SparkSQL) to enable you to get the most out of your data, and to deliver the insights that will drive your competitive advantage.

Who this book is for

This book is there to help anyone who wants to know more about using SQL to deliver analysis. This means that you could be:

- A data analyst
- A student
- A database developer
- An accountant
- A business analyst

Or indeed anyone who needs to deliver accurate analytics from the data stored in Databricks.

Why Databricks

Databricks is one of the world’s best-selling and widely used data repositories. It follows that learning the Databricks flavour of SQL (called SPARK SQL) and the basics of the Databricks database has the potential to be a career-enhancing move. As the data store for innumerable corporate databases, this mature and impressive system is used to power data analysis across the globe.

However, SPARK SQL is, fortunately, in proximity to the SQL used by rival databases. So learning *Databricks SQL* will set you on the path to applying SQL query techniques in most of the available SQL databases that are currently deployed.

What this book will bring you

This book was written to help the reader become proficient in querying databases using SQL. It is designed to help you to master a language that can seem arcane or even weird at first sight. To overcome any initial reticence that you may have, it progresses step by step through all the core concepts and techniques that you need to master. This way, you learn the essential keywords needed to query Spark SQL databases progressively without *information overload*.

To make SQL comprehensible we have chosen to introduce each new concept or keyword individually, so that you can learn each element in isolation. As the book progresses you see how SQL keywords can be combined to extend the power of SQL as you learn to create ever more powerful queries.

However no one queries databases purely for fun. So each query that you apply in this book also has a purely practical purpose. Therefore, you also see how to develop real-world queries that deliver essential data analysis. These queries can then be adapted to your own requirements using your own data.

How to read this book

This book has been designed to be read to use it in several different ways, depending on your knowledge of SQL and your real-world requirements.

SQL Novices

If you are a complete beginner, then you can begin with Chapter 1 and progress through the book until you feel that you have attained a level of SQL skills that match your needs. The book is designed to be a complete SQL querying course that allows readers with no previous SQL experience to gain the skills and experience that they need.

Refreshing your knowledge

If you are coming back to SQL after a while away, then you should probably skim through the first few chapters until you start meeting techniques and approaches that are less obvious. Then you can slow down and concentrate on progressing through the book and consolidating and refreshing your knowledge.

In-Depth querying

If your needs are more advanced, then you can probably skip Chapters 1 through 8 and start from Chapter 10. If necessary, you can use the next few chapters to provide inspiration on how best to solve your specific problems. You can then take a deep dive into Parts III and IV to make sure that your advanced SQL querying knowledge is up to mark.

Transferring SQL knowledge from another flavor of SQL

You may prefer to speed read the first few chapters to make sure that you are clear on the differences between SPARK SQL and the version of SQL that you have come from. You can then concentrate on progressing through the remainder of the book to acquire a thorough grounding in querying Spark SQL.

Solving specific issues

If you have been using SQL for a while but need some guidance to help you solve new kinds of query challenges then you can use this book as a collection of “recipes”.

In this case we suggest that you refer to the Solutions List that follows the table of contents. Here you can find most of the individual sections of this book grouped by their functional focus. You can then use the examples given as a basis for solving your own query requirements.

The structure of this book

The book is set out in three parts. Each corresponds to a level of SQL querying knowledge that you can use depending on your requirements.

Part I

This first part of the book is aimed at true Spark SQL beginners. It presumes no previous knowledge of SQL or Spark SQL and helps you progressively to acquire the core knowledge that is required to carry out basic SQL queries.

It consists of the following **nine** chapters:

Chapter 1: Writing Basic SQL Queries

Chapter 2: Filtering Data

Chapter 3: Applying Complex Filters to Queries

Chapter 4: Simple Calculations

Chapter 5: Aggregating Output

Chapter 6: Working with Dates in Databricks

Chapter 7: Formatting Text in Query Output

Chapter 8: Formatting Numbers and Dates

Chapter 9: Using Basic Logic to Enhance Analysis

Part II

The second part of this book concentrates on teaching you how to handle sets of data.

It consists of the following **seven** chapters:

Chapter 10: Using Multiple Tables When Querying Data

Chapter 11: Using Advanced Table Joins

Chapter 12: Subqueries

Chapter 13: Derived Tables

Chapter 14: Common Table Expressions

Chapter 15: Correlated Subqueries

Chapter 16: Datasets Manipulation

Part III

The third part of this book aims to extend your SQL knowledge so that you can use it to solve real-world analytical problems.

It consists of the following **five** chapters:

Chapter 17: Using SQL for More Advanced Calculations

Chapter 18: Segmenting and Classifying Data

Chapter 19: Rolling Analysis

Chapter 20: Analysing Data Over Time

Chapter 21: Complex Data Output

The sample data and sample queries

To help you learn SQL the sample data is available on the BPB website. This way you can practice SQL querying using the data that we have elaborated to help you learn SQL.

Code Bundle and Coloured Images

Please follow the link to download the *Code Bundle* and the *Coloured Images* of the book:

<https://rebrand.ly/yfha2jw>

The code bundle for the book is also hosted on GitHub at **<https://github.com/bpbpublications/Querying-Databricks-with-Spark-SQL>**. In case there's an update to the code, it will be updated on the existing GitHub repository.

We have code bundles from our rich catalogue of books and videos available at **<https://github.com/bpbpublications>**. Check them out!

Errata

We take immense pride in our work at BPB Publications and follow best practices to ensure the accuracy of our content to provide with an indulging reading experience to our subscribers. Our readers are our mirrors, and we use their inputs to reflect and improve upon human errors, if any, that may have occurred during the publishing processes involved. To let us maintain the quality and help us reach out to any readers who might be having difficulties due to any unforeseen errors, please write to us at :

errata@bpbonline.com

Your support, suggestions and feedbacks are highly appreciated by the BPB Publications' Family.

Did you know that BPB offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.bpbonline.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at :

business@bpbonline.com for more details.

At **www.bpbonline.com**, you can also read a collection of free technical articles, sign up for a range of free newsletters, and receive exclusive discounts and offers on BPB books and eBooks.

Piracy

If you come across any illegal copies of our works in any form on the internet, we would be grateful if you would provide us with the location address or website name. Please contact us at **business@bpbonline.com** with a link to the material.

If you are interested in becoming an author

If there is a topic that you have expertise in, and you are interested in either writing or contributing to a book, please visit **www.bpbonline.com**. We have worked with thousands of developers and tech professionals, just like you, to help them share their insights with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Reviews

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions. We at BPB can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about BPB, please visit **www.bpbonline.com**.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

<https://discord.bpbonline.com>



Table of Contents

1. Writing Basic SQL Queries.....	1
Introduction	1
Structure	2
Objectives	3
Databricks.....	3
The Databricks Web interface.....	3
Getting started with Databricks notebooks.....	5
Cells	7
Spark SQL.....	7
Databricks databases (schemas) and tables.....	7
Activating a database.....	9
Displaying the data in a table.....	9
<i>How does it work</i>	9
Help writing SQL	11
Limiting the number of records displayed.....	12
<i>How does it work</i>	13
Displaying data from a specific field.....	13
<i>How does it work</i>	14
Finding the columns in a table	15
<i>SQL writing style</i>	16
Displaying data from a specific set of fields.....	16
<i>How does it work</i>	17
<i>Columns or fields</i>	17
Modifying the field name in the output using aliases	17
<i>How does it work</i>	18
Removing duplicates from query output	19
<i>How does it work</i>	20
<i>Undo and redo</i>	20
Sorting data.....	20

<i>How does it work</i>	21
Sorting data in reverse order	22
<i>How does it work</i>	22
Applying multiple sort criteria	22
<i>How does it work</i>	23
Running SQL queries.....	24
Displaying the available tables.....	24
Finding all the views in a database.....	25
Resizing book cells.....	26
Overriding the 1,000-row output limit.....	26
Conclusion.....	26
2. Filtering Data.....	29
Introduction	29
Structure	29
Objectives	30
Filtering text.....	30
<i>How does it work</i>	31
Applying multiple text filters.....	32
<i>How does it work</i>	33
SQL writing style.....	33
Excluding an element	34
<i>How does it work</i>	34
Using multiple exclusion filters	35
<i>How does it work</i>	36
Filtering numbers over a defined threshold.....	36
<i>How does it work</i>	37
Filtering numbers under a defined threshold.....	37
<i>How does it work</i>	38
Filtering on values up to and including a specific number	38
<i>How does it work</i>	39
Filtering on a range of values.....	39
<i>How does it work</i>	40

Using Boolean filters (true or false)	41
<i>How does it work</i>	42
Conclusion.....	43
3. Applying Complex Filters to Queries.....	45
Introduction	45
Structure	45
Objectives	46
Using either/or filters.....	46
<i>How does it work</i>	47
Using multiple separate criteria concurrently	48
<i>How does it work</i>	48
Using multiple filters and an exclusion	49
<i>How does it work</i>	49
Filtering on both text and numbers simultaneously.....	50
<i>How does it work</i>	51
Applying complex alternative filters at the same time.....	52
<i>How does it work</i>	53
Removing case-sensitivity in filters.....	54
<i>How does it work</i>	55
Using wildcard searches.....	56
<i>How does it work</i>	56
Forcing case-insensitivity in wildcard filters	58
<i>How does it work</i>	59
Using wildcards to exclude data.....	59
<i>How does it work</i>	60
Applying alternative wildcard patterns	60
<i>How does it work</i>	61
Using a specific part of the text to filter data	62
<i>How does it work</i>	63
Filtering NULLs or nonexistent data.....	64
<i>How does it work</i>	65
Searching using regular expressions	66

<i>How does it work</i>	67
Conclusion.....	68
4. Simple Calculations	69
Introduction	69
Structure	69
Objectives	70
Doing simple math.....	70
<i>How does it work</i>	71
Examining data types in SQL tables and views.....	72
Isolating sections of formulas when applying math.....	73
<i>How does it work</i>	74
Calculating ratios	74
<i>How does it work</i>	75
Increasing values by a defined percentage.....	75
<i>How does it work</i>	76
Ordering the output of calculations	77
<i>How does it work</i>	77
NULLs	78
Handling missing data	79
<i>How does it work</i>	80
Filtering on a calculation.....	81
<i>How does it work</i>	81
Using complex calculated filters	82
<i>How does it work</i>	83
Operator precedence.....	84
Conclusion.....	85
5. Aggregating Output.....	87
Introduction	87
Structure	87
Objectives	88
Calculating table totals	88

<i>How does it work</i>	89
Using calculated aggregations	89
<i>How does it work</i>	90
Using grouped aggregations	91
<i>How does it work</i>	91
Using multiple levels of grouping	92
<i>How does it work</i>	93
Calculating averages	94
<i>How does it work</i>	94
Counting grouped elements	95
<i>How does it work</i>	95
Counting unique elements	96
<i>How does it work</i>	97
Displaying upper and lower numeric thresholds	97
<i>How does it work</i>	98
Aggregating across multiple columns	98
<i>How does it work</i>	99
Aggregating across columns and rows	99
<i>How does it work</i>	100
Filtering groups	101
<i>How does it work</i>	101
Filtering on aggregated results	102
<i>How does it work</i>	102
Selecting data based on aggregated results as well as specific filter criteria ...	103
<i>How does it work</i>	104
<i>Query analysis</i>	105
Sorting by aggregated results	105
<i>How does it work</i>	106
Finding elements where every Boolean value is true	106
<i>How does it work</i>	107
Conclusion	108

6. Working with Dates in Databricks.....	111
Introduction	111
Structure	111
Objectives	112
Dates in Databricks	112
<i>Filtering records by date</i>	<i>113</i>
<i>How does it work</i>	<i>113</i>
<i>Date datatypes in tables.....</i>	<i>115</i>
Using a range of dates to filter data.....	115
<i>How does it work</i>	<i>116</i>
<i>Finding the number of days between two dates.....</i>	<i>117</i>
<i>How does it work</i>	<i>117</i>
<i>Aggregating data over a date range.....</i>	<i>118</i>
<i>How does it work</i>	<i>119</i>
<i>Filtering by year</i>	<i>119</i>
<i>How does it work</i>	<i>120</i>
<i>Filtering records over a series of years.....</i>	<i>122</i>
<i>How does it work</i>	<i>122</i>
<i>Find sales for a specific day of the month</i>	<i>123</i>
<i>How does it work</i>	<i>124</i>
<i>Isolating data for a specific year and month</i>	<i>125</i>
<i>How does it work</i>	<i>125</i>
Finding data for a given quarter	126
<i>How does it work</i>	<i>126</i>
Filtering data by weekday	127
<i>How does it work</i>	<i>128</i>
Finding records for a specific week of the year	128
<i>How does it work</i>	<i>129</i>
Aggregating data by the day of the week in a given year.....	129
<i>How does it work</i>	<i>130</i>
Analyzing data by day of year.....	131
<i>How does it work</i>	<i>132</i>

Grouping data by the full weekday.....	133
<i>How does it work</i>	133
Displaying cumulative data over a period of months to a specific date.....	136
<i>How does it work</i>	136
Displaying cumulative data over 90 days up to a specific date	137
<i>How does it works</i>	138
Displaying the data for the previous three months	138
<i>How does it work</i>	139
Finding the current system date	140
Conclusion.....	140
7. Formatting Text in Query Output.....	143
Introduction	143
Structure	143
Objectives	144
Adding text to the output	144
<i>How does it work</i>	144
<i>Adding text to numbers</i>	145
<i>How does it work</i>	146
<i>Amalgamating columns</i>	146
<i>How does it work</i>	147
Avoiding NULLs in text-based data.....	147
<i>How does it work</i>	148
Concatenating and grouping.....	149
<i>How does it work</i>	149
<i>Adding multiple pieces of text to numbers</i>	150
<i>How does it work</i>	150
<i>Converting text to uppercase</i>	151
<i>How does it work</i>	151
<i>Converting text to lowercase</i>	152
<i>How does it work</i>	153
<i>Converting text to initial capitals</i>	153
Extracting the first few characters from a field.....	154

<i>How does it work</i>	154
<i>Displaying the three characters at the right of the text</i>	155
<i>How does it work</i>	156
<i>Displaying a given number of characters at a specific place in the text</i>	157
<i>How does it work</i>	157
<i>Replacing NULLs with the contents of another field</i>	158
<i>How does it work</i>	159
Filtering records based on the part of a field	160
<i>How does it work</i>	160
Filtering data using specific characters at a given position inside a field	161
<i>How does it work</i>	162
Conclusion	162
8. Formatting Numbers and Dates	165
Introduction	165
Structure	165
Objectives	166
Removing the decimals from the output	166
<i>How does it work</i>	167
Rounding a field up to the nearest whole number	167
<i>How does it work</i>	168
Rounding a value to the nearest whole number	168
<i>How does it work</i>	169
Rounding to a specific number of decimals	170
Rounding a value up or down to the nearest thousand	170
<i>How does it work</i>	171
Banker's rounding	172
Displaying a value in a specific numeric format	173
<i>How does it work</i>	173
Displaying a value in a specific currency	175
<i>How does it work</i>	176
Outputting a date in a specific date format	177
<i>How does it work</i>	177

Presenting the time in a specific format.....	179
Conclusion.....	181
9. Using Basic Logic to Enhance Analysis.....	183
<i>Introduction</i>	183
<i>Structure</i>	183
<i>Objectives</i>	184
Generating an alert when a value is too high	184
<i>How does it work</i>	185
Shortening text and adding ellipses to indicate truncation	186
<i>How does it work</i>	187
Designing complex calculated alerts.....	187
<i>How does it work</i>	188
Creating key performance indicators	190
<i>How does it work</i>	191
Classifying a series of elements without the necessary categories present in your data.....	192
<i>How does it work</i>	193
Creating ad hoc category groupings	195
<i>How does it work</i>	196
Applying multiple ad hoc categories	197
<i>How does it work</i>	198
Categorizing data using nested classifications	199
<i>How does it work</i>	200
Choosing elements from a list.....	201
<i>How does it work</i>	202
Placing nulls at the start or end of a list.....	204
<i>How does it work</i>	205
Conclusion.....	206
10. Using Multiple Tables When Querying Data.....	207
Introduction	207
Structure	207
Objectives	208

Joining tables.....	209
<i>How does it work</i>	211
<i>Other join syntax</i>	213
Joining multiple tables.....	214
<i>How does it work</i>	215
<i>Join fields</i>	217
Join subtleties.....	218
<i>How does it work</i>	219
Using table aliases.....	221
<i>How does it work</i>	221
Joining many tables.....	224
<i>How does it work</i>	226
Visualizing databases	227
Querying across databases.....	228
Conclusion.....	229
11. Using Advanced Table Joins	231
Introduction	231
Structure	231
Objectives	232
Filtering data using inner joins	232
Filtering data using multiple tables join.....	233
<i>How does it work</i>	234
Semi joins.....	235
<i>How does it work</i>	236
Filtering data output using intermediate tables	237
<i>How does it work</i>	238
Using left joins to return all the data in one table but not from the other table	239
<i>How does it work</i>	240
Right joins to return all the data in one table but not from the other.....	243
<i>How does it work</i>	244
Full joins to return all the data from both tables in a join	245

<i>How does it work</i>	246
Intermediate table joins	248
<i>How does it work</i>	249
Using multiple fields in joins	249
<i>How does it work</i>	250
Joining a table to itself	252
<i>How does it work</i>	253
Joining tables on ranges of values	254
<i>How does it work</i>	256
Cross joins	257
<i>How does it work</i>	258
Join concepts	259
Conclusion	259
12. Subqueries	261
Introduction	261
Structure	261
Objectives	262
Adding aggregated fields to detailed datasets	262
<i>How does it work</i>	263
Displaying a value as the percentage of a total	265
<i>How does it work</i>	266
Using a subquery to filter data	268
<i>How does it work</i>	268
Using a subquery as part of a calculation to filter data	269
<i>How does it work</i>	270
Filtering on an aggregated range of data using multiple subqueries	272
<i>How does it work</i>	272
Filtering on aggregated output using a second aggregation	273
<i>How does it work</i>	274
Using multiple results from a subquery to filter data	275
<i>How does it work</i>	276
Complex aggregated subqueries	278

<i>How does it work</i>	280
Nested subqueries.....	281
<i>How does it work</i>	282
Using subqueries to exclude data.....	283
<i>How does it work</i>	285
Filtering across queries and subqueries.....	286
<i>How does it work</i>	287
Applying separate filters to the Subquery and the main query	289
<i>How does it work</i>	290
Conclusion.....	291
13. Derived Tables	293
Introduction	293
Structure	293
Objectives	294
Using a derived table to create intermediate calculations	294
<i>How does it work</i>	296
Grouping and ordering data using a custom classification	299
<i>How does it work</i>	301
Joining derived tables with other tables	302
<i>How does it work</i>	303
Joining multiple derived tables.....	305
<i>How does it work</i>	307
Using multiple derived tables for complex aggregations	309
<i>How does it work</i>	311
<i>Data visibility</i>	313
Using derived tables to join unconnected tables	313
<i>How does it work</i>	314
Compare year-on-year data using a derived table.....	316
<i>How does it work</i>	317
Synchronizing filters between a derived table and the main query.....	318
<i>How does it work</i>	320
Conclusion.....	320

14. Common Table Expressions	323
Introduction	323
Structure	323
Objectives	324
Simplifying complex queries with CTEs	324
A basic common table expression	325
<i>How does it work</i>	326
Calculating averages across multiple values using a CTE.....	328
<i>How does it work</i>	329
<i>CTE or derived table?</i>	330
Reusing CTEs in a query	330
<i>How does it work</i>	331
Using a CTE in a derived table to deliver two different levels of aggregation	333
<i>How does it work</i>	335
Using a CTE to isolate data from a separate dataset at a different level of detail	336
<i>How does it work</i>	338
Multiple common table expressions in a single query	339
<i>How does it work</i>	341
Nested CTEs.....	342
<i>How does it work</i>	344
Using multiple CTEs to compare disparate datasets	346
<i>How does it work</i>	348
Conclusion.....	349
15. Correlated Subqueries.....	351
Introduction	351
Structure	351
Objectives	352
Simple correlated subqueries	352
<i>How does it work</i>	353
Correlated subqueries to display percentages of a specific total	355

<i>How does it work</i>	356
Comparing datasets using a correlated subquery	357
<i>How does it work</i>	358
Avoid correlated subqueries in certain cases	359
Duplicating the output of a correlated subquery in the query results	360
<i>How does it work</i>	361
Using correlated subqueries to filter data on an aggregate value	361
<i>How does it work</i>	362
<i>What makes a query correlated</i>	364
Using correlated subqueries to detect if records exist	364
<i>How does it work</i>	365
<i>Using a correlated subquery to exclude data</i>	366
<i>How does it work</i>	367
Conclusion	368
16. Datasets Manipulation	369
Introduction	369
Structure	369
Objectives	370
Read data from multiple identical tables using the UNION operator	370
<i>How does it work</i>	371
Isolate identical data in multiple tables using the INTERSECT operator	373
<i>How does it work</i>	374
Isolating nonidentical records using the EXCEPT operator	376
<i>How does it work</i>	377
Joining multiple identical tables in a subquery	378
<i>How does it work</i>	379
Conclusion	380
17. Using SQL for More Advanced Calculations	381
Introduction	381
Structure	381
Objectives	382

Calculating the percentage represented by each record in a dataset.....	383
<i>How does it work</i>	384
Replacing multiple subqueries.....	384
<i>How does it work</i>	385
Remove decimals in calculations.....	386
<i>How does it work</i>	387
Numeric datatypes.....	387
Converting between numeric datatypes.....	388
<i>How does it work</i>	388
Avoiding divide-by-zero errors.....	390
<i>How does it work</i>	391
Finding the remainder in a division using the modulo function.....	392
<i>How does it work</i>	393
Creating financial calculations	394
<i>How does it work</i>	396
Using a tally table to produce a sequential list of numbers.....	397
<i>How does it work</i>	398
Generating completely random sample output from a dataset	400
<i>How does it work</i>	400
Handling source data where figures are stored as text.....	401
<i>How does it work</i>	402
Conclusion.....	403
18. Segmenting and Classifying Data.....	405
Introduction	405
Structure	405
Objectives	406
Organizing data by rank	406
<i>How does it work</i>	407
Creating multiple groups of rankings.....	408
<i>How does it work</i>	409
Creating multiple ranked groups and subgroups.....	410
<i>How does it work</i>	412

Filtering data by ranked items	413
<i>How does it work</i>	414
Classifying data by strict order of rank.....	415
<i>How does it work</i>	417
Segment data into deciles.....	419
<i>How does it work</i>	420
Plot values for a percentile.....	421
<i>How does it work</i>	422
Extracting data from a specific quintile	423
<i>How does it work</i>	424
Display median values	425
<i>How does it work</i>	426
Conclusion.....	426
19. Rolling Analysis	429
Introduction	429
Structure	429
Objectives	430
Adding a running total.....	431
<i>How does it work</i>	431
Using windowing functions in an aggregated query	433
<i>How does it work</i>	434
Grouping running totals	434
<i>How does it work</i>	435
Applying windowing functions to a subquery	437
<i>How does it work</i>	438
Adding unique ids on the fly using ROW_NUMBER()	439
<i>How does it work</i>	440
Displaying records for missing data.....	442
<i>How does it work</i>	443
Displaying a complete range of dates and relevant data	445
<i>How does it work</i>	447
Comparing data with the data from a previous record.....	449

<i>How does it work</i>	449
Comparing data over time using the FIRST_VALUE() and LAST_VALUE() functions.....	452
<i>How does it work</i>	453
Displaying rolling averages over a specified number of records.....	455
<i>How does it work</i>	456
Show the first sale and last four sales per client.....	457
<i>How does it work</i>	458
Calculating cumulative distribution	459
<i>How does it work</i>	460
Classifying data using the PERCENT_RANK() function.....	461
<i>How does it work</i>	462
Using the LAG() function with alphabetical data	463
<i>How does it work</i>	464
Conclusion.....	464
20. Analyzing Data Over Time.....	467
Introduction	467
Structure	467
Objectives	468
Aggregating values for the year to date.....	468
<i>How does it work</i>	469
Isolating data for the previous month.....	471
<i>How does it work</i>	472
Using a derived table to compare data with values from a previous year.....	474
How does it work.....	475
Finding the total amount for sales each weekday over a year	476
<i>How does it work</i>	477
Count the number of weekend days between two dates.....	477
<i>How does it work</i>	478
Aggregate data for the last day of the month	479
<i>How does it work</i>	480
Aggregate data for the last Friday of the month	482

<i>How does it work</i>	483
Analyzing timespans as years, months, and days	485
<i>How does it work</i>	486
Isolate time periods from date and time data	488
<i>How does it work</i>	489
Listing data by time of day	490
<i>How does it work</i>	491
Aggregating data by hourly bandings.....	492
<i>How does it work</i>	492
Aggregate data by a quarter of an hour.....	493
<i>How does it work</i>	494
Reading dates and times stored as strings	494
<i>How does it work</i>	495
Conclusion.....	496
21. Complex Data Output.....	499
Introduction	499
Structure	499
Objectives	500
Creating a pivot table.....	500
<i>How does it work</i>	501
Creating a pivot table displaying multiple row groupings	504
<i>How does it work</i>	505
Adding totals to aggregate queries.....	507
<i>How does it work</i>	508
Creating subtotals and totals in aggregated queries.....	508
<i>How does it work</i>	509
Creating clear tables that include totals and subtotals	511
<i>How does it work</i>	514
Replace acronyms with full text in the final output.....	514
<i>How does it work</i>	515
Conclusion.....	516
Index	517-525

CHAPTER 1

Writing Basic SQL Queries

Introduction

Welcome to Databricks and the new world of data analysis that you are about to experience. As you are standing on the threshold of this voyage into the realm of serverless analytics, you could be feeling a little apprehensive. Well, do not worry; your journey will be as simple and comprehensible as possible. This chapter will start you on your adventure, first by outlining the software that you will be using and then by explaining what Databricks is. Then, it will show you how to take an initial look at the data itself. As you progress, you will learn how to be more selective about the data that you analyze.

It may seem obvious, but you will need some data in an accessible dataset before you can start your analysis. Throughout this book, you will be developing your analytical skills with the aid of a sample dataset named *Prestige Cars*. This dataset contains a small amount of data concerning sales of vehicles by a fictitious British car reseller. If you want to try the examples in this chapter, you will have to download the sample data from the BPB website and install it into a version of Databricks. So, it follows that now could be a good time to set up the sample dataset as described in *Appendix A* unless you have already done so. Of course, you can install the sample data only if Databricks is already available. So, if you are not in an enterprise environment

where Databricks is already accessible, you will need to set up an account for the Databricks community edition before anything else.

Structure

In this chapter, you will learn about the following topics:

- Databricks
- The Databricks Web interface
- Getting started with Databricks notebooks
- Cells
- Spark SQL
- Databricks databases (schemas) and tables
- Activating a database
- Displaying the data in a table
- Help writing SQL
- Limiting the number of records displayed
- Displaying data from a specific field
- Finding the columns in a table
- Displaying data from a specific set of fields
- Modifying the field name in the output
- Removing duplicates from query output
- Sorting data
- Sorting data in reverse order
- Applying multiple sort criteria
- Running SQL queries
- Displaying the available tables
- Finding all the views in a database
- Resizing book cells
- Overriding the 1,000-row output limit

Assuming that you have all the prerequisites in place, it is time to move on to the core focus of this chapter and start querying Databricks data.

Objectives

In this chapter, you will learn how to:

- Query Databricks databases using Databricks notebooks
- List the contents of tables
- Select only certain fields in tables to display
- Display only a few records from a table
- Give columns new names in your query output
- Sort your output

NOTE: If you know a little about SQL and if you have a basic knowledge of databases, then feel free to skip past the first few sections of this chapter (or even the first few chapters) until you find the parts that are new to you. If this is not the case, it is preferable to start from the beginning and provide all the information that you are likely to need to get the most out of your SQL learning experience.

Databricks

Databricks is a cloud data platform that delivers serverless analytics. This means that all the data and processing take place in the cloud (currently, Amazon Web Services, Microsoft Azure, or Google Cloud). All the analytics and data querying that you will do is carried out in a Web browser. Unlike traditional databases, Databricks does not load data into a database but stores data in files in a cloud-based data lake—hence the term *serverless analytics*.

Another area where Databricks breaks with the traditional database paradigm is by separating data storage from processing. This approach allows for much greater scalability and the ability to handle huge amounts of data—making Databricks a leader in the area of Big Data.

The technical approaches that underpin Databricks are extremely complex, but fortunately, the Databricks interface hides virtually all the complexity. All you have to do is load data (or access data that is already loaded) and start querying.

The Databricks Web interface

To start querying data in Databricks, you first need to log on. If you have a login to an enterprise version of Databricks, then you will have been given a corporate login

to use. If not (and assuming that you have created a community account), then all you need to do is to enter the following URL in your Web browser:

<https://community.cloud.databricks.com/login.html>

This will take you to the Databricks startup page, as shown in *Figure 1.1*:

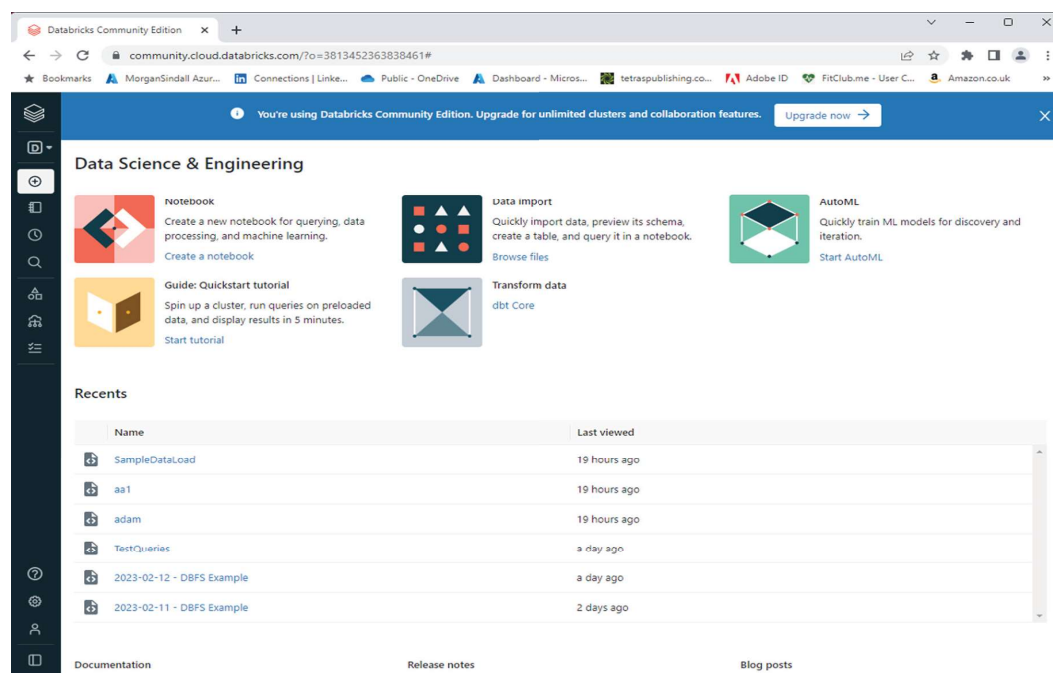


Figure 1.1: The Databricks startup page

Hover over the left sidebar to expand it. This is the main *menu* where you can find all the core Databricks options. You can see this in *Figure 1.2*:

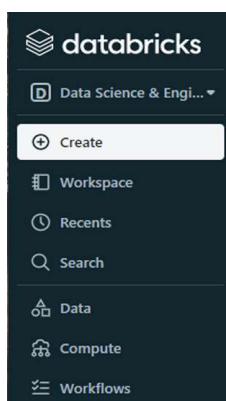


Figure 1.2: The Databricks main menu

NOTE: This book is only focused on one aspect of Databricks-querying data with Spark SQL. Consequently, there are huge swathes of Databricks that we will not be looking at. If you need to delve into other aspects of the platform, then the Databricks documentation is a good place to start.

Getting started with Databricks notebooks

Interaction with data in Databricks takes place in Databricks notebooks. You can consider these to be a kind of programming window where you will write your code to analyze and return data. The first thing to do is to create a new notebook. To create a new notebook, follow the following steps:

1. Hover over the plus symbol in the left pane and select **Notebook** in the popup pane. You can see this in *Figure 1.3*:

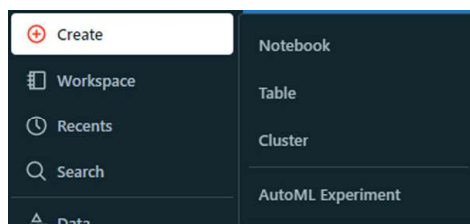


Figure 1.3: The Create pane

2. In the dialog that appears, enter a name for your notebook, ensure that the default language is set to SQL, and choose the cluster that you will be using to compute results. You can see this in *Figure 1.4*:

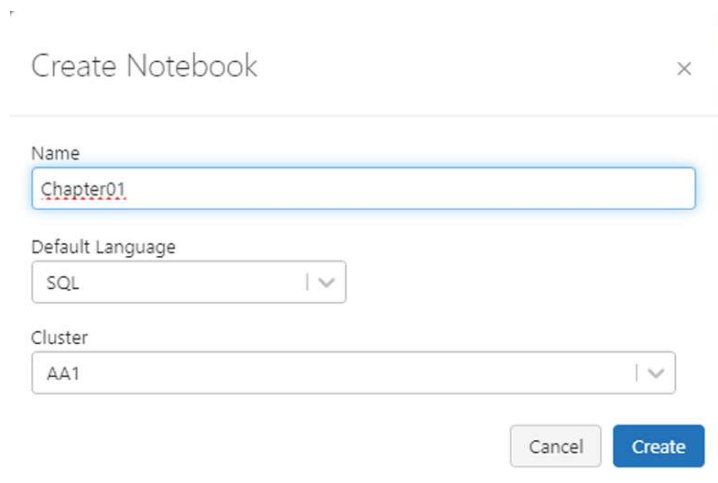


Figure 1.4: The Create Notebook dialog