Querying Databricks with Spark SQL

Leverage SQL to query and analyze Big Data for insights

Adam Aspin

www.bpbonline.com

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.

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

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 is 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*:

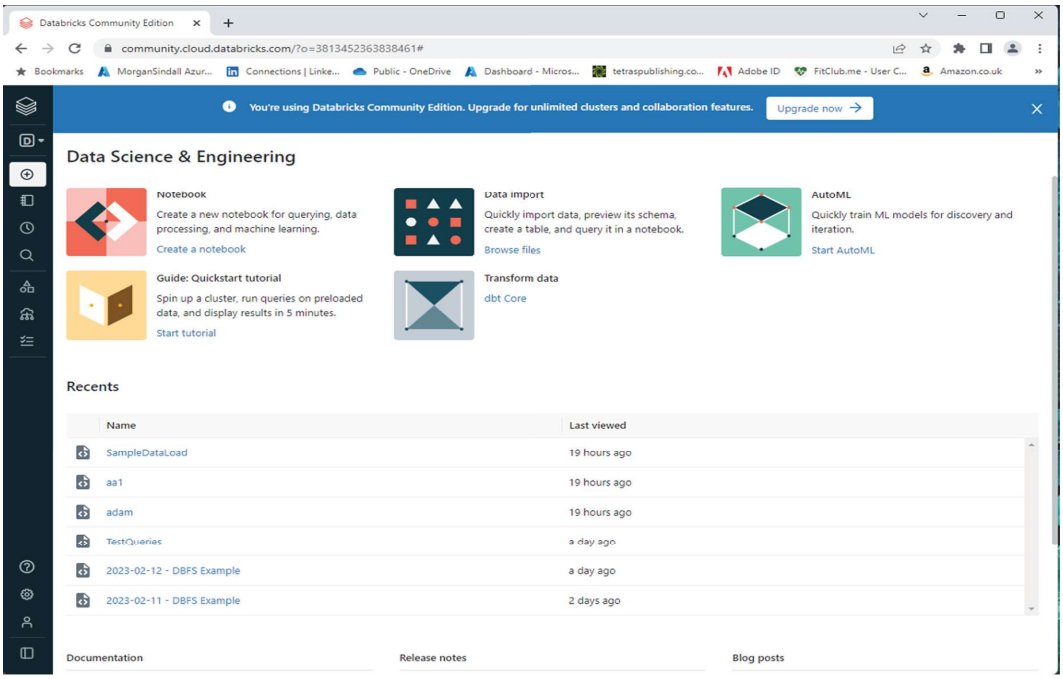

Figure 1.1: *The Databricks startup page*

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*:

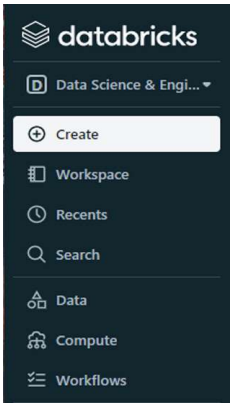

Figure 1.2: *The Databricks main menu*

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*:

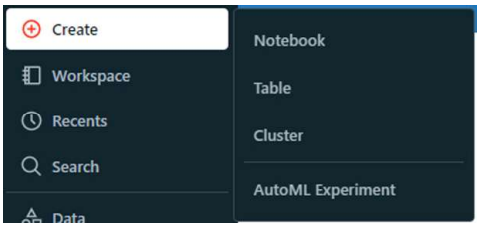

***Figure 1.3**: The Create pane*

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*:

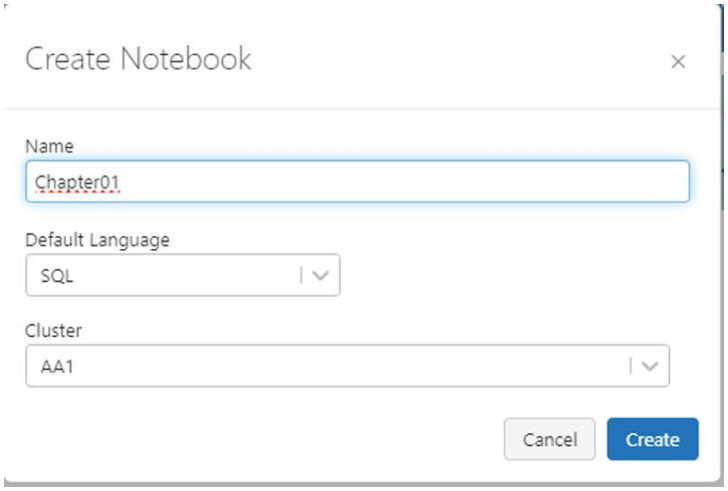

***Figure 1.4**: The Create Notebook dialog*