Data-Centric Systems and Applications

Antonio Badia

SQL for Data Science

Data Cleaning, Wrangling and Analytics with Relational Databases

Springer

Data-Centric Systems and Applications

Intelligent data management is the backbone of all information processing and has hence been one of the core topics in computer science from its very start. This series is intended to offer an international platform for the timely publication of all topics relevant to the development of data-centric systems and applications. All books show a strong practical or application relevance as well as a thorough scientific basis. They are therefore of particular interest to both researchers and professionals wishing to acquire detailed knowledge about concepts of which they need to make intelligent use when designing advanced solutions for their own problems.

Special emphasis is laid upon:

- Scientifically solid and detailed explanations of practically relevant concepts and techniques
  (what does it do)
- Detailed explanations of the practical relevance and importance of concepts and techniques
  (why do we need it)
- Detailed explanation of gaps between theory and practice
  (why it does not work)

According to this focus of the series, submissions of advanced textbooks or books for advanced professional use are encouraged; these should preferably be authored books or monographs, but coherently edited, multi-author books are also envisaged (e.g. for emerging topics). On the other hand, overly technical topics (like physical data access, data compression etc.), latest research results that still need validation through the research community, or mostly product-related information for practitioners ("how to use Oracle 9i efficiently") are not encouraged.

More information about this series at http://www.springer.com/series/5258

Antonio Badia

SQL for Data Science

Data Cleaning, Wrangling and Analytics with Relational Databases

Springer

Antonio Badia
Computer Engineering & Computer Science
University of Louisville
Louisville, KY, USA

ISSN 2197-9723 ISSN 2197-974X (electronic)
Data-Centric Systems and Applications
ISBN 978-3-030-57591-5 ISBN 978-3-030-57592-2 (eBook)
https://doi.org/10.1007/978-3-030-57592-2

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Data Science (or Data Analytics, or whatever one prefers to call it) is a 'hot' topic right now. There is an explosion of courses on the subject, especially online: many universities and several for-profit and non-profit organizations (Coursera, edX, Udacity, Udemy, DataCamp, and many others) offer on-campus and online courses, certification, and degrees. The coverage of these offerings is quite diverse, reflecting the fact that Data Science is still a young and evolving field. However, many courses seem to coalesce around a few topics (Machine Learning, mostly) and tools (R, Python, and SQL, mostly). What few of these courses offer is a textbook.

There are already many books on databases and SQL, but almost all of them focus on the traditional curriculum for Computer Science majors or Information Systems majors (there are a few exceptions, like [11] and [17]). In contrast, the present book explains SQL *within the context* of Data Science and is more in line with what is being taught in these new courses. This book introduces the different parts of SQL as they are needed for the tasks usually carried out during data analysis. Using the framework of the *data life cycle*, it focuses on the steps that are given the short shift in traditional textbooks, like data loading, cleaning, and pre-processing.

This book is for anyone interested in Data Science and/or databases. It should prove useful to anyone taking any of the abovementioned courses, online or on-campus, as well as to students working on their own. It assumes very little from the reader; it just demands a bit of 'computer fluency,' but no background on databases or data analysis. In general, all concepts are introduced intuitively and with a minimum of specialized jargon. It contains an appendix (Appendix A) meant to help students without prior experience with databases, with instructions on how to download and install the two open-source database systems (MySQL and Postgres) that we use for examples throughout the book. All readers of the book are encouraged to install both systems and follow the book along with a computer in order to practice, do the exercises, and play around—simply reading the book alone is going to be much less useful than *using* it.

The book is organized as follows: Chapter 1 describes the *Data Life Cycle*, the sequence of stages, from data acquisition and ingestion until archiving, that data goes through as it is prepped for analysis and then actually analyzed, together with

the different activities that take place at each stage. It also explains the different ways that datasets can be organized, and the different types of data one may have to deal with. Many students have an intuitive understanding of the concepts in this chapter, but it is useful to have it all together in one place and to give a name to each concept for later reference. Chapter 2 gets into databases proper, explaining how relational databases organize data. The chapter also explains how data in tables *should* look like (what Hadley Wickham has called *tidy* data [19]), a point which is not traditionally emphasized and can lead to severe problems down the road. Non-traditional data, like XML and text, are also covered. Chapter 3 introduces SQL *queries*, the SQL commands that allow us to ask questions about the data. Unlike traditional textbooks, queries and their parts are described around typical data analysis tasks (data exploration, cleaning, and transformation). These tasks are vital for a proper examination of the data but are frequently overlooked in Data Mining and Machine Learning textbooks. Chapter 4 introduces some basic techniques for Data Analysis. Even though this is not the focus of the book, the chapter shows that SQL can be used for some simple analyses without too much complication.

After this part, which constitutes the core of the book, Chap. 5 introduces additional SQL constructs that come in handy in a variety of situations. This chapter completes the coverage of SQL queries so that readers get an overview of all the main aspects of this important topic. Chapter 6 briefly explains how to use SQL from within R and from within Python programs. This chapter is not an introduction to R (or to Python) and, unlike other chapters in the book, does assume that the reader is already familiar with at least the basics of R and Python. It focuses on how these languages can interact with a database, and how what has been learned about SQL can be leveraged to make life easier when using R or Python.

The book also contains another appendix (besides the one already mentioned), which introduces some basic approaches for handling very large datasets. The purpose of this appendix is to demystify the ideas behind the vague label *Big Data* and give the readers basic guidance on how to use their newly acquired skills in this world.

As in many textbooks, none of what this one contains is new. This book covers the same (or very similar) content to what can be found in many sources, especially online. What this book does is to put it all together under one roof and to give it some order and structure. In many blogs and sites, the material is presented as an answer to a particular question (how do you…?), which may be useful to someone with a specific need but gives the impression that learning SQL is about a bag of tricks. Here, the material is logically organized using the idea of the data life cycle so that all the concepts introduced can be understood as parts of a coherent whole.

Data Science itself is a relatively new and still changing field, but it has deep roots, as it uses approaches and techniques from well-established fields, mostly math (statistics, linear algebra, and others) and computer science (databases, machine learning, and others). As a result, the same concept is sometimes given different names by different authors in different textbooks. Whenever I am aware of this, I

have given a list of known names so that readers with different backgrounds can relate what is in here with what they already know.

The goal of the book is to introduce some basic concepts to a wide variety of readers and provide them a good foundation on which they can build. After going through this book, readers should be able to profitably learn more about Data Mining, Machine Learning, and database management from more advanced textbooks and courses. It is my hope that most of them feel that they have been given a springboard from which they are in a good position to dive deeper into the fascinating world of data analysis.

Louisville, KY, USA Antonio Badia
July 2020

Contents