Preface

We live in a world where the amount of data being generated is constantly increasing. While a few decades ago, an organization may have had a single database that could store everything they needed to track, today most organizations have tens, hundreds, or even thousands of databases, along with data warehouses, and perhaps a data lake. And these data stores are being fed from an increasing number of data sources (transaction data, web server log files, IoT and other sensors, and social media, to name just a few).

It is no surprise that we hear more and more companies talk about being data-driven in their decision making. But in order for an organization to be truly data-driven, they need to be masters of managing and drawing insights from these ever-increasing quantities and types of data. And to enable this, organizations need to employ people with specialized data skills.

Doing a search on LinkedIn for jobs related to data returns nearly 800,000 results (and that is just for the United States!). The job titles include roles such as data engineer, data scientist, and data architect.

This revised edition of the book includes updates to all chapters, covering new features and services from AWS, as well as three brand-new chapters. In these new chapters, we cover topics such as building transactional data lakes (using open table formats such as Apache Iceberg), implementing a data mesh approach on AWS, and using a DataOps approach to building a modern data platform.

While this book will not magically turn you into a data engineer, it has been designed to accelerate your journey toward data engineering on AWS. By the end of this book, you will not only have learned some of the core concepts around data engineering, but you will also have a good understanding of the wide variety of tools available in AWS for working with data. You will also have been through numerous hands-on exercises, and thus gained practical experience with things such as ingesting streaming data, transforming and optimizing data, building visualizations, and even drawing insights from data using AI.

Who this book is for

This book has been designed for two groups of people; firstly, those looking to get started with a career in data engineering, and who want to learn core data engineering concepts. This book introduces many different aspects of data engineering, providing a comprehensive high-level understanding of, and practical hands-on experience with, different focus areas of data engineering.

Secondly, this book is for those people who may already have an established career focused on data, but who are new to the cloud, and to AWS in particular. For these people, this book provides a clear understanding of many of the different AWS services for working with data, and gives them hands-on experience with a variety of these AWS services.

What this book covers

Each of the chapters in this book takes the approach of introducing important concepts or key AWS services, and then providing a hands-on exercise related to the topic of the chapter:

Chapter 1, An Introduction to Data Engineering, reviews the challenges of ever-increasing dataset volumes, and the role of the data engineer in working with data in the cloud.

Chapter 2, *Data Management Architectures for Analytics*, introduces foundational concepts and technologies related to big data processing.

Chapter 3, *The AWS Data Engineer's Toolkit*, provides an introduction to a wide range of AWS services that are used for ingesting, processing, and consuming data, and orchestrating pipelines.

Chapter 4, *Data Governance, Security, and Cataloging*, covers the all-important topics of keeping data secure, ensuring good data governance, and the importance of cataloging your data.

Chapter 5, *Architecting Data Engineering Pipelines*, provides an approach for whiteboarding the high-level design of a data engineering pipeline.

Chapter 6, *Ingesting Batch and Streaming Data*, looks at the variety of data sources that we may need to ingest from, and examines AWS services for ingesting both batch and streaming data.

Chapter 7, *Transforming Data to Optimize for Analytics*, covers common transformations for optimizing datasets and for applying business logic.

Chapter 8, *Identifying and Enabling Data Consumers*, is about better understanding the different types of data consumers that a data engineer may work to prepare data for.

Chapter 9, *A Deeper Dive into Data Marts and Amazon Redshift*, focuses on the use of data warehouses as a data mart and looks at moving data between a data lake and data warehouse. This chapter also does a deep dive into Amazon Redshift, a cloud-based data warehouse.

Chapter 10, *Orchestrating the Data Pipeline*, looks at how various data engineering tasks and transformations can be put together in a data pipeline, and how these can be run and managed with pipeline orchestration tools such as AWS Step Functions.

Chapter 11, *Ad Hoc Queries with Amazon Athena*, does a deeper dive into the Amazon Athena service, which can be used to run SQL queries directly on data in the data lake, and beyond.

Chapter 12, *Visualizing Data with Amazon QuickSight*, discusses the importance of being able to craft visualizations of data, and how the Amazon QuickSight service enables this.

Chapter 13, *Enabling Artificial Intelligence and Machine Learning*, reviews how AI and ML are increasingly important for gaining new value from data, and introduces some of the AWS services for both ML and AI.

Chapter 14, *Building Transactional Data Lakes*, looks at new table formats (including Apache Iceberg, Apache Hudi, and Delta Lake) that bring traditional data warehousing type features to data lakes.

Chapter 15, *Implementing a Data Mesh Strategy*, discusses a recent trend, referred to as a data mesh, that provides a new way to approach analytical data management and data sharing within an organization.

Chapter 16, *Building a Modern Data Platform on AWS*, introduces important concepts, such as DataOps, which provides automation and observability when building a modern data platform.

Chapter 17, *Wrapping Up the First Part of Your Learning Journey*, concludes the book by looking at the bigger picture of data analytics, including real-world examples of data pipelines, and a review of emerging trends in the industry.

To get the most out of this book

Basic knowledge of computer systems and concepts, and how these are used within large organizations, is helpful prerequisite knowledge for this book. However, no data engineering-specific skills or knowledge are required. Also, a familiarity with cloud computing fundamentals and core AWS systems will make it easier to follow along, especially with the hands-on exercises, but detailed step-by-step instructions are included for each task.

> **Note:**
>
> If you are using the digital version of this book, we advise you to access the code from the book's GitHub repository (a link is available in the next section), rather than copying and pasting from the PDF or electronic version. Doing so will help you avoid any potential formatting errors when copying and pasting code.

Download the example code files

The code bundle for the book is hosted on GitHub at **`https://github.com/PacktPublishing/Data-Engineering-with-AWS-2nd-edition`**. We also have other code bundles from our rich catalog of books and videos available at **`https://github.com/PacktPublishing/`**. Check them out!

Download the color images

We also provide a PDF file that has color images of the screenshots/diagrams used in this book. You can download it here: **`https://packt.link/gbp/9781804614426`**.

Conventions used

There are a number of text conventions used throughout this book.

`CodeInText` : Indicates code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles. For example: "Include a `WHERE Year = 2020` clause."

A block of code is set as follows:

```
datalake_bucket/year=2023/file1.parquet
datalake_bucket/year=2022/file1.parquet
datalake_bucket/year=2021/file1.parquet
datalake_bucket/year=2020/file1.parquet
```

When we wish to draw your attention to a particular part of a code block, the relevant lines or items are set in bold:

```
datalake_bucket/year=2023/file1.parquet
datalake_bucket/year=2022/file1.parquet
datalake_bucket/year=2021/file1.parquet
datalake_bucket/year=2020/file1.parquet
```

Bold: Indicates a new term, an important word, or words that you see on the screen. For instance, words in menus or dialog boxes appear in the text like this. For example: "In addition, you can use **Spark SQL** to process data using standard SQL."

Warnings or important notes appear like this.

Tips and tricks appear like this.

Get in touch

Feedback from our readers is always welcome.

General feedback: Email `feedback@packtpub.com` and mention the book's title in the subject of your message. If you have questions about any aspect of this book, please email us at `questions@packtpub.com`.

Errata: Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you have found a mistake in this book, we would be grateful if you reported this to us. Please visit `http://www.packtpub.com/submit-errata`, click **Submit Errata**, and fill in the form.

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 `copyright@packtpub.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 `http://authors.packtpub.com`.

Share your thoughts

Once you've read *Data Engineering with AWS, Second Edition*, we'd love to hear your thoughts! Please **`click here to go straight to the Amazon review page`** for this book and share your feedback.

Your review is important to us and the tech community and will help us make sure we're delivering excellent quality content.

Download a free PDF copy of this book

Thanks for purchasing this book!

Do you like to read on the go but are unable to carry your print books everywhere? Is your eBook purchase not compatible with the device of your choice?

Don't worry, now with every Packt book you get a DRM-free PDF version of that book at no cost.

Read anywhere, any place, on any device. Search, copy, and paste code from your favorite technical books directly into your application.

The perks don’t stop there, you can get exclusive access to discounts, newsletters, and great free content in your inbox daily

Follow these simple steps to get the benefits:

1. Scan the QR code or visit the link below

`https://packt.link/free-ebook/9781804614426`

2. Submit your proof of purchase
3. That’s it! We’ll send your free PDF and other benefits to your email directly