The Definitive Guide to Azure Data Engineering

Modern ELT, DevOps, and Analytics on the Azure Cloud Platform

Ron C. L'Esteve

Apress®

The Definitive Guide to Azure Data Engineering: Modern ELT, DevOps, and Analytics on the Azure Cloud Platform

Ron C. L'Esteve
Chicago, IL, USA

ISBN-13 (pbk): 978-1-4842-7181-0 ISBN-13 (electronic): 978-1-4842-7182-7
https://doi.org/10.1007/978-1-4842-7182-7

Copyright © 2021 by Ron C. L'Esteve

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

Trademarked names, logos, and images may appear in this book. Rather than use a trademark symbol with every occurrence of a trademarked name, logo, or image we use the names, logos, and images only in an editorial fashion and to the benefit of the trademark owner, with no intention of infringement of the trademark.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Jonathan Gennick
Development Editor: Laura Berendson
Coordinating Editor: Jill Balzano

Cover image designed by Freepik (www.freepik.com)

Distributed to the book trade worldwide by Springer Science+Business Media LLC, 1 New York Plaza, Suite 4600, New York, NY 10004. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail booktranslations@springernature.com; for reprint, paperback, or audio rights, please e-mail bookpermissions@springernature.com.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at http://www.apress.com/bulk-sales.

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub via the book's product page, located at www.apress.com/9781484271810. For more detailed information, please visit http://www.apress.com/source-code.

Printed on acid-free paper

For Mom and Dad.

Table of Contents

About the Author xvii

About the Technical Reviewer xix

Acknowledgments xxi

Introduction xxiii

Part I: Getting Started 1

Chapter 1: The Tools and Prerequisites 3

Master the Traditional Microsoft Business Intelligence Stack 3

Understand Azure's Modern Enterprise Data and Analytics Platform 4

Understand How to Manage Big Data with Azure 6

Understand the Fundamental Requirements for the Data Engineer Associate 7

Expand Your Knowledge Across Azure Specialties 8

Be Able to Address the Business Value of the Azure Data Platform 10

Get Hands-On with Azure Data Engineering Through Azure Portal 11

Azure Services Covered in This Book 11

Data Lake Storage Gen2 12

Data Factory 12

Databricks 15

Synapse Analytics 15

DevOps CI/CD 16

IoT Hub 16

Stream Analytics 17

Power BI 18

Purview 20

Snowflake 21

SQL Database 21

Cosmos DB 23

Relevant Azure Services Not Covered 25

Analysis Services 25

Cognitive Services 26

Azure Machine Learning 28

Monitor 29

Log Analytics 29

Event Hubs 29

Data Share 29

Logic Apps 29

Power Apps 30

App Service 30

SQL Managed Instance 30

Data Box 30

Data Sync 31

Data Gateway 31

Cost Management + Billing 31

Digital Twins 31

Mobile 31

Networking 31

Security 32

Identity 32

Kubernetes 32

Functions 32

HVR Real-Time Data Replication 32

Summary 33

Chapter 2: Data Factory vs. SSIS vs. Databricks 35

Choosing the Right Data Integration Tool 35

When to Use Azure Data Factory, Azure Databricks, or Both 37

Summary 39

Chapter 3: Design a Data Lake Storage Gen2 Account 41
Data Lake Layers 41
Environments.. 42
Storage Accounts 43
File Systems 46
Zones, Directories, and Files 48
Security 52
Control Plane Permissions... 52
Data Plane Permissions.. 52
POSIX-Like Access Control Lists.. 53
Shared Access Signature.. 53
Data Encryption 54
Network Transport 54
Summary... 54
Part II: Azure Data Factory for ELT...55
Chapter 4: Dynamically Load a SQL Database to Data Lake Storage Gen2............ 57
Azure Prerequisite Resources 58
Prepare and Verify SQL Server Database Objects 60
Prepare and Verify Azure SQL Database Objects 62
Prepare an Azure Data Lake Storage Gen2 Container... 63
Create Azure Data Factory Pipeline Resources 64
Create a Self-Hosted Integration Runtime.. 65
Create Linked Services... 66
Create Datasets 67
Create Azure Data Factory Pipelines... 71
P_Insert_Base_Table_Info... 72
P_SQL_to_ADLS 75
Run the Data Factory Pipeline and Verify Azure Data Lake Storage Gen2 Objects 79
Summary... 80

Chapter 5: Use COPY INTO to Load a Synapse Analytics Dedicated SQL Pool........ 83
Features of the COPY INTO Command.. 84
Data Preparation Tips... 85
Tip #1: Remove Spaces from the Column Names.. 85
Tip #2: Convert VARCHAR(MAX) to VARCHAR(4000)... 86
COPY INTO Using a Parquet File.. 86
COPY INTO Using a CSV File... 87
Using COPY INTO from Data Factory.. 89
Summary.. 91
Chapter 6: Load Data Lake Storage Gen2 Files into a Synapse Analytics Dedicated SQL Pool.. 93
Recreate the Pipeline Parameter Table... 94
Create the Datasets... 96
DS_ADLS_TO_SYNAPSE.. 96
DS_ADLS_TO_SYNAPSE_MI... 99
DS_SYNAPSE_ANALYTICS_DW.. 101
Create the Pipeline.. 102
Choose the Copy Method... 106
BULK INSERT.. 106
PolyBase... 109
Copy Command... 111
Summary.. 113
Chapter 7: Create and Load Synapse Analytics Dedicated SQL Pool Tables Dynamically.. 115
Dynamically Create and Load New Tables Using an ADF Pre-copy Script............. 116
Dynamically Truncate and Load Existing Tables Using an ADF Pre-copy Script...... 124
Dynamically Drop, Create, and Load Tables Using a Stored Procedure.................. 125
Summary.. 130

Chapter 8: Build Custom Logs in SQL Database for Pipeline Activity Metrics...... 133

Option 1: Create a Stored Procedure Activity 134

Option 2: Create a CSV Log File in Data Lake Storage Gen2 137

Option 3: Create a Log Table in Azure SQL Database 143

Summary........ 146

Chapter 9: Capture Pipeline Error Logs in SQL Database 147

Create a Parameter Table........ 148

Create a Log Table........ 150

Create an Errors Table........ 151

Create a Stored Procedure to Update the Log Table 152

Create a Stored Procedure to Update the Errors Table 155

Create a Source Error........ 158

Add Records to a Parameter Table........ 159

Verify the Azure Data Lake Storage Gen2 Folders and Files 160

Configure the Pipeline Lookup Activity 160

Configure the Pipeline ForEach Loop Activity 161

Configure a Stored Procedure to Update the Log Table 161

Configure a Stored Procedure to Update the Errors Table 165

Run the Pipeline........ 167

Verify the Results........ 168

Other ADF Logging Options 170

Summary........ 172

Chapter 10: Dynamically Load a Snowflake Data Warehouse 173

Linked Services and Datasets........ 174

Base Linked Services 174

Datasets 180

Snowflake Control Database and Tables........ 185

Pipelines 187

Step 1: Design and Execute an ADF Pipeline to Load Azure SQL Database to Data Lake Storage Gen2 187

Step 2: Design the Data Lake Storage Gen2 to Snowflake ADF Pipeline 196

Comparing the Various ADLS Gen2 to Snowflake Ingestion Options 211

Swim Lanes 212

Data Validation 213

Summary 214

Chapter 11: Mapping Data Flows for Data Warehouse ETL 215

Modern Data Warehouse 216

Creating the Base Azure Data Resources 216

Slowly Changing Dimension Type I 218

Create a Data Factory Pipeline and Datasets 219

Create a Data Factory Mapping Data Flow 221

Exists 225

LookupDates 225

SetAttributes 226

AlterRows 226

sink1 227

Updating a Record 230

Inserting a Record 233

Summary 236

Chapter 12: Aggregate and Transform Big Data Using Mapping Data Flows 237

Add Files and Folders to Azure Data Lake Storage Gen2 237

File Size 238

Folder Structure 238

Create Azure Data Factory Resources 240

Create the Mapping Data Flow 242

Regular Expressions (Regex) 247

Soundex 248

RANK Function.. 251

DENSE_RANK Function... 252

ROW_NUMBER Function... 253

Summary... 256

Chapter 13: Incrementally Upsert Data 259

Create a Parameter Table.. 260

Create a Source Query for the ADF Pipeline.. 262

Add the ADF Datasets 263

Azure SQL Database 264

Azure Data Lake Storage Gen2.. 264

Azure Synapse Analytics DW 266

Create the ADF Pipeline 267

Add a Lookup Activity to Get the List of Tables.. 267

Add a ForEach Activity to Iterate and Copy Each Table.. 268

Mapping Data Flow for SQL to Lake Incremental ADF Pipeline 269

Mapping Data Flow to Incrementally Upsert from Lake to Synapse Analytics DW............. 278

Run the ADF Pipeline.. 285

Verify Incremental SQL to Lake Pipeline Results.. 285

Verify Incremental Upsert Lake to Synapse ADF Pipeline Results.............................. 287

Verify Source SQL Record Count 287

Verify Lake Folder and Parquet File Path... 288

Verify Destination Synapse Record Count 288

Insert a Source SQL Record... 289

Verify Incremental SQL to Lake ADF Pipeline Results... 290

Verify Incremental Upsert Lake to Synapse ADF Pipeline Results.............................. 290

Verify Destination Synapse Analytics DW Record Count.. 291

Update a Source SQL Record.. 292

Verify Destination Synapse Analytics DW Record Count.. 292

Summary... 294

Chapter 14: Load Excel Sheets into Azure SQL Database Tables........................ 295

Prerequisites.. 295

Create an Excel Spreadsheet.. 296

Upload to Azure Data Lake Storage Gen2... 297

Create Linked Services and Datasets... 297

Create a Pipeline to Load Multiple Excel Sheets in a Spreadsheet into a Single Azure SQL Table 302

Create a Pipeline to Load Multiple Excel Sheets in a Spreadsheet into Multiple Azure SQL Tables.. 308

Summary.. 320

Chapter 15: Delta Lake 321

Why an ACID Delta Lake.. 321

Prerequisites.. 322

Create and Insert into Delta Lake.. 325

Update Delta Lake... 333

Delete from Delta Lake 339

Explore Delta Logs 343

Insert 344

Update 344

Delete 345

Summary.. 346

Part III: Real-Time Analytics in Azure...347

Chapter 16: Stream Analytics Anomaly Detection.. 349

Prerequisites.. 349

Create an Azure Stream Analytics Job... 350

Create an IoT Hub 352

Create a Power BI Service 358

Download the Device Simulator 358

Create a Stream Analytics Input and Output.. 360

Add Stream Input.. 360

Add Stream Output 364

Write the Stream Analytics Query ... 368

Start the Stream Analytics Job ... 369

Create a Real-Time Power BI Dashboard ... 371

Create a Dataset ... 372

Create a Dashboard ... 372

Add a Tile ... 373

Run the Device Simulator ... 377

Monitor Real-Time Power BI Streaming ... 379

Summary ... 381

Chapter 17: Real-Time IoT Analytics Using Apache Spark ... 383

Prerequisites ... 384

Create an IoT Hub ... 385

Create a Databricks Cluster ... 386

Install Maven Library ... 386

Create a Notebook and Run Structured Streaming Queries ... 388

Configure Notebook Connections ... 388

Start the Structured Stream ... 391

Start the IoT Device Simulator ... 393

Display the Real-Time Streaming Data ... 394

Create a Spark SQL Table ... 395

Write the Stream to a Delta Table ... 396

Summary ... 399

Chapter 18: Azure Synapse Link for Cosmos DB ... 401

Create an Azure Cosmos DB Account ... 402

Enable Azure Synapse Link ... 404

Create a Cosmos DB Container and Database ... 406

Import Data into Azure Cosmos DB ... 409

Create a Cosmos DB Linked Service in Azure Synapse Analytics ... 413

Load and Query the Data Using Synapse Spark ... 415

Summary ... 419

Part IV: DevOps for Continuous Integration and Deployment 421

Chapter 19: Deploy Data Factory Changes 423

Prerequisites 423

Create the DevOps Continuous Integration Build Pipeline 431

Create the DevOps Continuous Deployment Release Pipeline 437

Azure PowerShell Task to Stop Triggers 441

ARM Template Deployment Task 442

Azure PowerShell Task to Start Triggers 445

Run the Release Pipeline 446

Verify the Deployed Data Factory Resources 448

Summary 453

Chapter 20: Deploy a SQL Database 455

Pre-Requisites 456

Create a Visual Studio SQL Database Project 457

Install Visual Studio GitHub Extension 458

Import AdventureWorks Database 459

Connect to GitHub Repo Source Control 462

Check In Visual Studio Solution to GitHub Repo 464

Install Azure Pipelines from GitHub 467

Build CI Pipeline from GitHub Repo 472

Release CD Pipeline from DevOps Artifact Repo 483

Verify Deployed Azure SQL AdventureWorks Database 491

Summary 493

Part V: Advanced Analytics 495

Chapter 21: Graph Analytics Using Apache Spark's GraphFrame API 497

Install JAR Library 498

Load New Data Tables 500

Load Data in a Databricks Notebook 504

Build a Graph with Vertices and Edges 504

Query the Graph 506
Find Patterns with Motifs 508
Discover Importance with PageRank 510
Explore In-Degree and Out-Degree Metrics 512
Run a Breadth-First Search 516
Find Connected Components 516
Summary 519
Chapter 22: Synapse Analytics Workspaces 521
Create a Synapse Analytics Workspace 522
Explore Sample Data with Spark 525
Query Data with SQL 533
Create External Table with SQL 536
Summary 541
Chapter 23: Machine Learning in Databricks 543
Create an MLflow Experiment 544
Install the MLflow Library 546
Create a Notebook 547
Selective Logging 547
Auto-logging 552
Register a Model 554
Summary 559
Part VI: Data Governance 561
Chapter 24: Purview for Data Governance 563
Create Azure Purview Account 564
Explore Azure Purview 569
Create and Register Data Source 570
Manage Credentials and Access 574
Create a Scan 578

Explore the Glossary 587
Browse Assets.. 589
Working with Purview Programmatically.. 595
Summary.. 598

Index.. 601

About the Author

Ron C. L'Esteve is a professional author residing in Chicago, IL, USA. His passion for Azure Data Engineering originates from his deep experience with designing, implementing, and delivering modern Azure data projects for numerous clients. Ron is a trusted technology leader and digital innovation strategist, responsible for scaling key data architectures, defining the road map and strategy for the future of data and business intelligence (BI) needs, and challenging customers to grow by thoroughly understanding the fluid business opportunities and enabling change by translating them into high-quality and sustainable technical solutions that solve the most complex challenges and promote digital innovation and transformation. He applies a practical and business-oriented approach of taking transformational ideas from concept to scale. Ron is an advocate for data excellence across industries and consulting practices and empowers self-service data, BI, and AI through his contributions to the Microsoft technical community.

About the Technical Reviewer

Greg Low is one of the better-known database consultants in the world. In addition to deep technical skills, Greg has experience with business and project management and is known for his pragmatic approach to solving issues. His skill levels at dealing with complex situations and his intricate knowledge of the industry have seen him cut through difficult problems.

Microsoft has specifically recognized his capabilities and appointed him to the Regional Director program. They describe it as consisting of "150 of the world's top technology visionaries chosen specifically for their proven cross-platform expertise, community leadership, and commitment to business results."

Greg leads a boutique data consultancy firm called SQL Down Under. His clients range from large tier 1 organizations to start-ups.

Greg is a long-term Data Platform MVP and considered one of the foremost consultants in the world on SQL Server and Microsoft data-related technologies. He has provided architectural guidance for some of the largest SQL Server implementations in the world and helped them to resolve complex issues. Greg was one of the two people first appointed as SQL Server Masters worldwide. Microsoft use him to train their own staff.

For several years, Greg served on the global board for the Professional Association for SQL Server. He is particularly proud of having helped it triple the size of its community and, more importantly to him, taken it from being 90% US based to being a truly global community with 60% of chapters outside the United States.

A talented trainer and presenter, Greg is known for his ability to explain complex concepts with great clarity to people of all skill levels. He is regularly invited to present at top-level tier 1 conferences around the world. Greg's SQL Down Under podcast has a regular audience of over 40,000 listeners.

Outside of work and family, Greg's current main passion is learning Mandarin Chinese, and he is determined to learn to read, write, speak, and understand it clearly.

Acknowledgments

Writing this book has been both a solitary and accompanied journey with sacrifices and victories along the way. Thank you to all who have supported me on the path to completing this book.

Introduction

With the numerous cloud computing technologies being at the forefront of the modern-day data architectural and engineering platforms, Microsoft Azure's cloud platform has contributed over 200 products and services that have been specifically designed to solve complex data challenges, empower self-service data engineering, and pave the way for the future of data and AI.

Navigating through these many offerings in the Azure Data Platform can become daunting for aspiring Azure Data Engineers, architects, consultants, and organizations that are seeking to build scalable, performant, and production-ready data solutions. This book is intended to uncover many of the complexities within the Azure data ecosystem with ease through structured end-to-end scenario-based demonstrations, exercises, and reusable architectural patterns for working with data in Azure and building highly performant data ingestion and ELT pipelines.

As Azure continues to introduce numerous data services to their ever-growing and evolving platform, this book will demystify many of the complexities of Azure Data Engineering with ease and introduce you to tried, tested, and production-ready patterns and pipelines for data of all different volumes, varieties, and velocities.

Additionally, you will be introduced to the many capabilities of bringing value and insights to your data through real-time and advanced analytics, continuously integrating and deploying your data ingestion pipelines, and getting started with many Azure data services to help you progress through your journey within the Azure Data Engineering ecosystem.