Mapping Data Flows in Azure Data Factory

Building Scalable ETL Projects in the Microsoft Cloud

—

Mark Kromer

Apress®

Mapping Data Flows in Azure Data Factory

Building Scalable ETL Projects in the Microsoft Cloud

Mark Kromer

Apress®

Mapping Data Flows in Azure Data Factory: Building Scalable ETL Projects in the Microsoft Cloud

Mark Kromer
SNOHOMISH, WA, USA

ISBN-13 (pbk): 978-1-4842-8611-1 ISBN-13 (electronic): 978-1-4842-8612-8
https://doi.org/10.1007/978-1-4842-8612-8

Copyright © 2022 by Mark Kromer

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 designed by eStudioCalamar

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

Distributed to the book trade worldwide by Springer Science+Business Media New York, 1 New York Plaza, Suite 4600, New York, NY 10004-1562, USA. 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 (https://github.com/Apress). For more detailed information, please visit http://www.apress.com/source-code.

Printed on acid-free paper

This book is dedicated to my loving wife Stacy and
our boys Ethan and Jude. Thank you for putting up with
my late hours working on data analytics and writing this book!

Table of Contents

About the Author xi

About the Technical Reviewer xiii

Introduction xv

Part I: Getting Started with Azure Data Factory and Mapping Data Flows 1

Chapter 1: ETL for the Cloud Data Engineer 3

General ETL Process 3

Differences in Cloud-Based ETL 5

Data Drift 8

Landing the Refined Data 9

Typical SDLC 10

Summary 12

Chapter 2: Introduction to Azure Data Factory 13

What Is Azure Data Factory? 13

Factory Resources 15

Pipelines 15

Activities 15

Triggers 15

Mapping Data Flows 16

Linked Services 16

Datasets 16

Azure Integration Runtime 16

Self-Hosted Integration Runtime 17

Elements of a Pipeline ... 18
Pipeline Execution ... 24
Pipeline Triggers ... 24
Pipeline Monitoring ... 25
Summary ... 26

Chapter 3: Introduction to Mapping Data Flows ... 27
Getting Started ... 27
Design Surface ... 30
Connector Lines and Reference Lines ... 31
Repositioning Nodes ... 32
Data Flow Script ... 35
Transformation Primitives ... 37
Multiple Inputs/Outputs ... 39
Schema Modifier ... 40
Formatters ... 42
Row Modifier ... 43
Flowlets ... 43
Destination ... 44
Expression language ... 44
Functions ... 44
Input Schema ... 44
Parameters ... 45
Cached Lookup ... 45
Locals ... 45
Data Preview ... 45
Manage Compute Environment from Azure IR ... 46
Debugging from the Data Flow Surface ... 48
Debugging from Pipeline ... 50
Summary ... 50

Part II: Designing Scalable ETL Jobs with ADF Mapping Data Flows 51

Chapter 4: Build Your First ETL Pipeline in ADF 53

Scenario 53

Data Quality 54

Task 1: Start with a New Data Flow 55

Task 2: Metadata Checker 57

Task 3: Add Asserts for Data Validation 58

Task 4: Filter Out NULLs 60

Task 5: Create Full Address Field 61

Final Step: Land the Data As Parquet in the Data Lake 63

Summary 65

Chapter 5: Common ETL Pipeline Practices in ADF with Mapping Data Flows 67

Task 1: Create a New Pipeline 67

Task 2: Debug the Pipeline 69

Task 3: Evaluate Execution Plan 71

Task 4: Evaluate Results 76

Task 5: Prepare Pipeline for Operational Deployment 77

Summary 78

Chapter 6: Slowly Changing Dimensions 79

Building a Slowly Changing Dimension Pattern in Mapping Data Flows 79

Data Sources 80

NewProducts 81

ExistingProducts 83

Cached Lookup 84

Create Cache 84

Create Row Hashes 84

Surrogate Key Generation 85

Check for Existing Dimension Members 85

Set Dimension Properties 87

Bring the Streams Together 89
Prepare Data for Writing to Database 89
Summary 92

Chapter 7: Data Deduplication 93
The Need for Data Deduplication 93
Type 1: Distinct Rows 94
Type 2: Fuzzy Matching 98
Column Pattern Matching 100
Self-Join 102
Match Scoring 106
Scoring Your Data for Duplication Evaluation 106
Turn the Data Flow into a Reusable Flowlet 109
Debugging a Flowlet 111
Summary 115

Chapter 8: Mapping Data Flow Advanced Topics 117
Working with Complex Data Types 117
Hierarchical Structures 118
Arrays 127
Maps 130
Data Lake File Formats 131
Parquet 131
Delta Lake 132
Optimized Row Columnar 132
Avro 133
JSON and Delimited Text 133
Data Flow Script 133
Summary 136

Part III: Operationalize Your ETL Data Pipelines 137

Chapter 9: Basics of CI/CD and Pipeline Scheduling 139

Configure Git 139

New Factory 140

Existing Factory 143

Branching 145

Publish Changes 147

Pipeline Scheduling 150

Debug Run 150

Trigger Now 151

Schedule Trigger 151

Tumbling Window Trigger 151

Storage Events Trigger 152

Custom Events Trigger 152

Summary 153

Chapter 10: Monitor, Manage, and Optimize 155

Monitoring Your Jobs 155

Error Row Handling 160

Partitioning Strategies 163

Optimizing Integration Runtimes 165

Compute Settings 165

Time to Live (TTL) 166

Iterating over Files 167

Parameterizing 167

Pipeline Parameters 168

Data Flow Parameters 172

Late Binding 174

Data Profiling 175

Mapping Data Flow Statistics 175

Data Preview Statistics 175

Profile Stats 176

Power Query Activity 177

Transformation Optimization 180

byName( ) and byNames( ) 180

Rank and Surrogate Key 181

Sorting 182

Database Queries 182

Joins and Lookups 183

Pipeline Optimizations for Data Flow Activity 185

Run in Parallel 186

Logging Level 187

Database Staging 187

Summary 187

Index 189