

# Contents

<b>Introduction</b>	<b>xxv</b>
<b>Chapter 1: SSIS Solution Architecture</b>	<b>1</b>
<b>Problem</b>	<b>2</b>
Macro Challenge: Enterprise Data Spaghetti	2
Micro Challenge: Data-Processing Confusion	4
Problems with Execution and Troubleshooting	5
Infrastructure Challenges	7
Other Challenges	8
<b>Design</b>	<b>8</b>
Choosing the Right Tool	9
Overall Solution Architecture	10
Data Integration or Consolidation	11
Data Warehouse ETL	12
Project Planning and Team Development	14
Agile Benefits	14
Agile Cautions and Planning	14
Data Element Documentation	15
Package Design Patterns	18
Modular Packages	18
Master Packages	19
Server and Storage Hardware	21
Server Hardware	21
Development and Test Servers	22
ETL Collocation	22
Storage Hardware	22
Package Execution Location	24
Package Storage Location Versus Execution Location	24
Execute SQL Task and Bulk Insert Task Execution	24
Package Execution and the Data Flow	25
Design Review	28
<b>Solution</b>	<b>28</b>
Setting the Stage: Management and Deployment	29
Source Data: Files, Tables, and Data Cleansing	29
Data Warehouse ETL and Cube Processing	30
Advanced ETL: Scripting, High Availability, and Performance	31
<b>Summary</b>	<b>31</b>

# Contents

---

## **Chapter 2: SSIS Management Framework Design** **33**

---

<b>Problem</b>	<b>33</b>
Challenges of Not Having a Framework Implementation	34
Different Development Methods	34
Changing Metadata	34
Getting the Right Information to the Right People at the Right Time	35
Reusability	35
Framework Implementation	35
Configuration Management Scheme	35
Logging and Auditing	36
Template Package	36
Framework Benefits and Value	36
<b>Design</b>	<b>37</b>
Configuration Management	38
Overall Design	38
Environment Variable Configuration	38
XML Configuration File	39
SQL Server Configuration	39
Logging and Auditing Mechanism	40
Storage	40
Tables	40
Stored Procedures	41
Template Package	42
Implementation Guidelines	43
<b>Solution</b>	<b>43</b>
Configuration Management	43
Logging and Auditing Mechanism	48
Storage and Tables	48
Stored Procedures	52
Template Package	69
Development	69
Installation	71
Other Considerations	71
Customizations	71
ETL Process Framework	72
Process Owner	72
Reporting	72
<b>Summary</b>	<b>73</b>

## **Chapter 3: Package Deployment and Storage Decisions** **75**

---

<b>Problem</b>	<b>75</b>
Standardization	76
Environment	76

---

Application	77
Desired Implementation	77
<b>Design</b>	<b>77</b>
Storage Methods	77
SSIS Service	78
File System Storage	79
SQL Server Storage	81
Storage Comparison	81
Deployment Methods	83
SSIS Deployment Wizard	83
Manual Deployment of Packages	88
DTUtil Scripting	90
<b>Solution</b>	<b>90</b>
Storage Methodology	91
Existing Infrastructure	92
Security	92
Package Makeup	93
Back to the Flowchart	93
Deployment Methodology	94
Storage Method	95
Existing Infrastructure	95
Deployment Needs	95
Back to the Flowchart	96
Total Methodology	96
<b>Summary</b>	<b>96</b>
<b>Chapter 4: File-Handling and Processing Methods</b>	<b>97</b>
<b>Problem</b>	<b>98</b>
Simple File Operations	98
Remote File Operations	98
File Criteria	98
File Iteration	98
File Notification	99
<b>Design</b>	<b>99</b>
File System Task	99
Source and Destination Connection	99
Attributes	100
Overwriting the Destination	101
Multiple Files	101
FTP Task	101
FTP Connection	102
HTTP	103
Foreach Loop	104
Applying Advanced Criteria	106

# Contents

---

File Watcher Task	109
WMI Event Watcher Task	110
<b>Solution</b>	<b>110</b>
<b>Summary</b>	<b>118</b>
<b>Chapter 5: Data Extraction Best Practices</b>	<b>119</b>
<b>Problem</b>	<b>120</b>
Extraction Data Criteria	120
Source System Impact	121
Incremental Extraction	121
Deleted Rows	122
Staging Database	123
Data Lineage and Metadata	123
File Sources	124
<b>Design</b>	<b>124</b>
Package Connections and Source Adapters	124
Package Connections	124
Source Adapters	128
Incremental Data Extraction	133
Incremental Extraction Using a Change Identifier Value	134
Targeting Change Records through Dynamic Source Queries	134
Retrieving Incremental Identifier Values and Updating Package Variables	137
Capturing the Maximum Change Identifier Value	141
Incremental Extraction from SQL Server without a Trigger	144
Using SSIS to Handle All Aspects of an Incremental Extraction	147
Data Staging Method	152
Tracking Data Lineage Identifiers	154
<b>Solution</b>	<b>156</b>
Metadata-Driven Extraction Example	156
Metadata Tables	157
Control Flow	157
Looping through the Sources	158
Inside the Sources Loop	160
The Script	161
Read Variables	162
Open Connections	162
Get Max Change Indicator	163
Extract Changed Source Rows	164
Close Connections	164
<b>Summary</b>	<b>165</b>

---

<b>Chapter 6: Data-Cleansing Design</b>	<b>167</b>
<b>Problem</b>	<b>168</b>
Candidate Key Profiles	170
Column Length Distribution Profiles	171
Column Null Ratio Profiles	172
Column Pattern Profiles	172
Column Statistics Profiles	173
Column Value Distribution Profiles	173
Functional Dependency Profiles	174
<b>Design</b>	<b>175</b>
Using the Script Transform for Data Scrubbing	175
Using the Fuzzy Grouping Transform to De-duplicate Data	177
Using the Fuzzy Lookup Transform to Cleanse Data	180
Dealing with Multiple Record Types in a Single File	183
Using the Raw File	188
<b>Solution</b>	<b>190</b>
<b>Summary</b>	<b>195</b>
<b>Chapter 7: Dimension Table ETL</b>	<b>197</b>
<b>Problem — Fundamental Dimension ETL</b>	<b>197</b>
Dimensions: The Basics	198
Dimension ETL: The Challenge	202
<b>Design — Fundamental Dimension ETL</b>	<b>203</b>
Data Preparation	203
Dimension Change Types	204
Type 1 (Changing) Attribute: A Closer Look	205
Type 2 (Historical) Attribute: A Closer Look	206
Inferred Members	207
<b>Solution — Fundamental Dimension ETL</b>	<b>209</b>
Preparing Your Source Data for Dimension ETL	210
SSIS Slowly Changing Dimension Wizard	214
Advanced Properties and Additional Outputs of the SCD	225
<b>Problem — Advanced Dimension ETL</b>	<b>227</b>
SCD Wizard Advantages and Disadvantages	227
Dimension Volume and Complexity	228
<b>Design — Advanced Dimension ETL</b>	<b>229</b>
Optimizing the Built-in SCD	229
Index Optimizations	229
Update Optimizations	229
Snowflake Dimension Tables	231
Parent-Child Dimension ETL	231

# Contents

---

Date Dimension ETL	233
Profile Dimension and Junk Dimension ETL	234
Creating a Custom Slowly Changing Package	234
<b>Solution — Advanced Dimension ETL</b>	<b>236</b>
Snowflake Dimensions	237
Parent-Child Dimensions	238
Profile and Junk Dimensions	240
Date Dimension ETL	243
Custom Dimension ETL	246
Determining Dimension Changes	248
Inserts and Updates	250
<b>Summary</b>	<b>252</b>
<b>Chapter 8: Fact Table ETL</b>	<b>253</b>
<b>Problem</b>	<b>253</b>
Fact Tables: The Basics	254
Fact Table ETL: The Challenge	257
Preparing the Data	258
Mapping Dimension Keys	258
Calculating Measures	258
Adding Metadata	259
Fact Table Updates	259
Fact Table Inserts	259
<b>Design</b>	<b>260</b>
Data Preparation	260
Data Preparation with SSIS Transformations	260
Data Preparation Examples	262
Acquiring the Dimension Surrogate Key in SSIS	265
Identifying the Dimension Surrogate Keys with SSIS	265
Surrogate Key Examples in SSIS	265
Measure Calculations	271
Measure Calculation Types	271
Handling Measure Calculations in SSIS	271
Managing Fact Table Changes	273
Approaches to Identifying Changed Fact Records	273
Fact Update Examples in SSIS	274
Optimizing Fact Table Inserts	278
Optimizing Inserts with Fast Load	279
Optimizing Inserts with Index Management	279
<b>Solution</b>	<b>280</b>
Internet and Reseller Sales Fact Table ETL	280
Fact Internet and Reseller Sales Extraction and Transform Process	281
Fact Internet and Reseller Sales Load Process	285

---

Snapshot Fact Table Example — Call Center Fact Table	289
Advanced Fact Table ETL Concepts	295
Handling Missing Dimension Lookups	295
Handling Late-Arriving Facts	300
<b>Summary</b>	<b>303</b>
<b>Chapter 9: SSAS Processing Architecture</b>	<b>305</b>
<b>Problem</b>	<b>305</b>
SSAS Object Population	306
Schedule	306
Partition Management	306
The Complete Package	307
<b>Design</b>	<b>307</b>
SSAS Objects and Processing Basics	307
Dimensions	308
Partitions	309
Mining Structures	311
SSIS Tasks and Components	312
Analysis Services Processing Task	313
Analysis Services Execute DDL Task	316
Execute Process Task with ASCMD	317
Data Flow Destinations for SSAS Objects	319
Script Task with AMO	321
Creating and Managing Partitions	323
Overall Design	323
<b>Solution</b>	<b>323</b>
Preparation for SSAS Integration	324
Process Dimensions Package	325
Process Task	325
Parallel XMLA	326
Process Partitions Package	327
Storing and Loading Metadata	328
SSAS Processing	329
Overall Solution	332
<b>Summary</b>	<b>333</b>
<b>Chapter 10: Implementing Scale-Out ETL Process</b>	<b>335</b>
<b>Problem</b>	<b>336</b>
<b>Design</b>	<b>337</b>
Design Components Overview	338
Central Common Services Server	338
File Processor and Pre-Aggregation Scale-Out Processes Servers	339

# Contents

---

Design Details	341
File Management Tasks	341
Data File Management	341
Work Allocation Process	342
Scale-Out Source File Process	343
Work Reassignment Process	343
Data Aggregation Tasks	344
Hourly Data Pre-Aggregation Process	344
Hourly Data Aggregation Process	345
Daily Data Aggregation Process	345
Archival and Clean-up Processes	345
Data File Archival Process	345
Stage Table Clean-up Process	346
Design Conclusions	346
<b>Solution</b>	<b>346</b>
Central Server Services	347
Multiple File Processor and Pre-Aggregator Processes	347
Database Tables Required on the Central Server	347
Stored Procedures	349
Procedures on the Central Server	349
Procedures on the Staging Servers	351
SSIS Packages	351
File-Processing Server Packages	351
Central Server Packages	354
<b>Summary</b>	<b>357</b>
<b>Chapter 11: Scripting Design Patterns</b>	<b>359</b>
<b>Problem — Advanced File Management</b>	<b>359</b>
Script Task	360
Scenario	360
<b>Design — Advanced File Management</b>	<b>360</b>
Script Language	361
Accessing Objects	362
Custom Assemblies	362
Scripting	363
<b>Solution — Advanced File Management</b>	<b>363</b>
Create an External Assembly	363
Access the External Assembly	366
Archive the Files	367
Summary — Advanced File Management	369
<b>Problem — Call Center Fact ETL</b>	<b>370</b>
Reasons to Use Scripting	370
Scenario	370

<b>Design — Call Center Fact ETL</b>	<b>371</b>
Component Type	371
Output Type	372
Synchronous Outputs	372
Asynchronous Outputs	373
Other Component Types	373
Design Choice	373
Overridable Methods	373
<b>Solution — Call Center Fact ETL</b>	<b>374</b>
Package Setup	375
Script Component	376
Reusable Methods	379
Row-By-Row Processing	380
Process Input	383
Entire Script	384
Package Wrap-Up	385
Summary — Call Center Fact ETL	387
<b>Summary</b>	<b>387</b>

---

**Chapter 12: SSIS Package Scaling** **389**

<b>Problem</b>	<b>390</b>
Identifying Task Durations	390
Identifying Data Flow Destination Issues	392
Identifying Transformation and Memory Bottlenecks	394
Identifying Data Flow Source Performance Issues	398
<b>Design</b>	<b>399</b>
Balancing SQL Operations with the Data Flow	399
Data Flow Advantages as Compared with SQL	401
SQL Advantages when Compared with the Data Flow	402
Applying SQL Code in Your SSIS Packages	402
SSIS Pipeline Architecture Concepts	403
Data Flow Design Example	404
SQL-Centric SSIS Process	404
Rewritten Data Flow–Centric SSIS Process	406
<b>Solution</b>	<b>407</b>
Tuning Your Data Flow	407
Use Blocking Transformations Sparingly	408
Limit Row-by-Row Operations	408
Manage Data Flow Backpressure	408
Pre-sort Sources as an Alternative to the Sort Transformation	408
Optimize the Lookup and Managing the Cache Size	409
Remove Unused Columns from the Data Flow	410
Be Mindful of the SSIS Logging Impact	411
Regulate Transactions	411

# Contents

---

Setting Data Flow Properties	411
Up the EngineThreads Value	412
Optimize the Temporary Storage Locations	412
Leave RunInOptimizedMode as True	413
Tune Buffers	413
Database and Destination Optimization	413
Limiting Database Growth	414
Consider Dropping and Re-creating Indexes on Destination Tables	414
Using the OLE DB Destination Adapter	414
Use Advanced Oracle and Teradata Destination Adapters	418
Handling Partitioned Tables	419
<b>Summary</b>	<b>422</b>
<b>Index</b>	<b>423</b>