Brought to you by:

Qlik LEAD WITH DATA | Microsoft

Data Warehouse Automation in Azure®

for dummies®
A Wiley Brand

Learn about data warehouse automation

Discover the benefits of Azure Synapse Analytics

Work with real-time data

Qlik Data Integration and Azure Synapse Special Edition

Matthew Basile
Clive Bearman
Rajeev Jain
Kevin Pardue

About Qlik

Qlik's vision is a data-literate world, one where everyone can use data to improve decision-making and solve their most challenging problems. Only Qlik offers end-to-end, real-time data integration and analytics solutions that help organizations access and transform all their data into value. Qlik's data integration solution, acquired from Attunity, lets you automate and stream data from a variety of sources to the Microsoft Data Platform quickly, easily, and securely. As an Azure Migration Partner, Qlik's data integration solutions are available in the Azure Marketplace or directly from Qlik. Qlik does business in more than 100 countries and serves over 50,000 customers around the world.
www.qlik.com

About Microsoft

Microsoft develops, manufactures, licenses, supports, and sells computer software, consumer electronics, personal computers, and related services. For more information, go to **www.microsoft.com.**

Data Warehouse Automation in Azure

Data Warehouse Automation in Azure

Qlik Data Integration and
Azure Synapse Special Edition

by

Matthew Basile

Clive Bearman

Rajeev Jain

Kevin Pardue

for dummies®

A Wiley Brand

Data Warehouse Automation in Azure For Dummies®, Qlik Data Integration and Azure Synapse Special Edition

Published by **John Wiley & Sons, Inc. 111 River St. Hoboken, NJ 07030-5774**

www.wiley.com

ISBN 978-1-119-62091-4 (pbk); ISBN 978-1-119-62090-7 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Publisher's Acknowledgments

Some of the people who helped bring this book to market include the following:

Project Editor: Rebecca Senninger

Editorial Manager: Rev Mengle

Business Development Representative: Susan Blessing

Production Editor: Mohammed Zafar Ali

Table of Contents

Introduction

Every business has data coming in from multiple sources and needs to move it to a data warehouse where it can be stored properly or used across the organization. Today's data warehouses are moving from on-premise to the cloud. This creates unique challenges to consider from their on-premise counterparts that you're probably already familiar with.

About This Book

This book gives you an introduction to data warehousing in Azure. You find out about its advantages in cost, time to value, and real-time data for analytics.

Icons Used in This Book


REMEMBER

Remember icons mark the information that's especially important to know.


TIP

The Tip icon points out helpful suggestions and useful nuggets of information.


WARNING

The Warning icon marks important information that may save you headaches.


TECHNICAL STUFF

If you're technically inclined, be sure to read the paragraphs marked with the Technical Stuff icon. If you don't geek out on the technical specs, skip 'em.

Beyond the Book

Azure Synapse Analytics is a fully managed cloud data warehouse for enterprises of any size that combines lightning-fast query performance with industry-leading data security. To learn more, go to: `https://azure.microsoft.com/en-us/services/synapse-analytics`.

Where to Go from Here

The book is modular so you can start with Chapter 1, but feel free to roam around to the chapters that fit best. You can also pick out a topic that interests you from the Table of Contents.

IN THIS CHAPTER

- Discovering the benefits of a cloud data warehouse
- Checking out Azure Synapse
- Things to consider when building your enterprise data warehouse

Chapter 1
What's Azure Synapse Analytics?

Every data pipeline has two major parts: ingestion/orchestration, and storage/compute. In this chapter, we start by talking about the storage/compute first — in this case, a data warehouse. We walk through the benefits of a cloud data warehouse over its on-premises counterparts. Then we dive more deeply into Azure Synapse, including its architecture, use-cases, and features. Finally, we cover some of the most important areas to consider when building your data warehouse — from securing your data to maximizing your spend.

Why a Cloud Data Warehouse?

Data warehouses have existed on-premises for decades, from Teradata and Netezza to Microsoft's own SQL Server. Over the years, the goals of the data warehouse haven't truly changed: consolidate data for performant analytics and reporting. In other words — they allow you to get value from your data. When comparing cloud data warehouses (such as Azure Synapse) to on-premise counterparts, there are many familiar concepts. Many of these warehouses use elastic scale-out architecture, store data in

a primarily relational format (tables), and use some flavor of SQL to query the data. Data is ingested from a variety of sources, and analysts either query the data warehouse directly, or through a reporting and business intelligence (BI) tool.

Yet, there are also many differences between an on-premise data warehouse and a cloud data warehouse. The most fundamental difference is that cloud data warehouses are, for the most part, Data Warehouses as a Service (DWaaS). This means some cloud provider (such as Azure) takes on the complexity of managing the underlying infrastructure. From lowering management cost and complexity to enabling scale and flexibility, cloud data warehouses and the DWaaS model enables patterns and architectures that were previously impossible on-premise while still fulfilling (and exceeding) a company's need for reporting and analytics. Over the next few sections, we walk through these benefits and use cases.

"AS A SERVICE"

The term "<blank> as a Service" refers to anything managed by a company but made available to a customer in an on-demand model, typically in the context of cloud computing. Three of the most common service models are IaaS, PaaS, and SaaS:

- **Infrastructure-as-a-Service:** Manages the physical hardware, the networking, and the servers. You're responsible for the OS, development tools, data, and applications. **Example:** Azure Virtual Machines.
- **Platform-as-a-Service**: Manages all the infrastructure and the middleware, development tools, data, and so on. You build the application on top. **Example:** Azure App Service, Azure Functions.
- **Software-as-a-Service:** Fully managed offering. The least customization, but the fastest and easiest to get started. **Example:** Microsoft IoT Central, Azure Machine Learning.

However, really any technology can exist "as a Service." In this case, Azure Synapse is a Data Warehouse as a Service — Azure handles the hardware and technology stack, while you provide the data and maintain control over the schema, security, and more.

Elasticity and scale

Perhaps the greatest advantage of cloud computing in general is its elasticity and flexibility, and this remains true for the cloud data warehouse. Previously, a company would have to build its infrastructure, which could range from a server room in the office to a data center, depending on the scale. This meant that even small-scale data projects required significant capital investment, creating a high barrier of entry for companies looking to build their data estate. And between planning, product selection, testing, and implementation, database projects were measured in months to years.

Cloud providers such as Azure have completely changed this model. Now, any company, regardless of size or age, can spin up a single virtual machine (VM) or tap into thousands of machines to run massive workloads, often in seconds or minutes. Combined with the massive increase in data (a topic for another book entirely), cloud computing has allowed both small startups and massive enterprises to upgrade and modernize their data estate without the pains of building on-premise architecture.

REMEMBER

While cloud data warehouses undoubtedly have major benefits, some problems can arise due to the differences in architecture and design. Here are some high-level things to look for when choosing a cloud data warehouse:

- **How long does it take to scale?** Some data warehouses scale in seconds, while others can take hours.
- **Do I get dedicated resources?** Some cloud architectures share resources across customers, which can affect consistency and compliance.
- **Can I scale compute and storage separately?** Some data warehouses are closer to hosted appliances and limit your flexibility like an on-premises appliance.
- **What are my cost controls?** You can always throw more hardware at a problem with the cloud, but it's essential you have tools to control your spend and maximize your investment.

Stress-free management

Passing off the infrastructure and hardware of your data warehouse to a cloud provider gives you more than just increased elasticity and scale. While it varies by cloud provider and data warehouse, cloud data warehouses often automatically handle many common management tasks for you, making your data warehouse even simpler to manage.

One common management task is ensuring your system has High Availability and Disaster Recovery (HADR), ensuring your data is always protected and available. With on-premises systems, this meant copying data into a secondary database that had to be built and installed in a different location. Installing the backup hardware, keeping the secondary database up to date, and manually restoring the data makes HADR a massive headache that only grows with your amount of data.

DWaaS offerings can simplify HADR drastically by either making it easy to spin up additional databases in separate regions, or by handling it automatically. For example, Azure Synapse automatically snapshots your data, lets you take manual snapshots, and provides automatic geo-backups as well. So, if you have sensitive data that must be replicated in multiple regions to ensure availability, this is a must-have feature when looking at which DWaaS to choose.

WARNING

Don't fire your DBAs yet! While DWaaS solutions may offset much of the management pain, remember that you must still manage your data warehouse. The cloud provider handles the infrastructure, but you're always responsible for classic database design tasks such as security models and data modeling.

Looking under the Hood of Azure Synapse

Now that we've established what makes cloud data warehouses special, we dive deep into Azure Synapse specifically. As Azure's first-party cloud data warehouse offering, Azure Synapse is a central part of any big data pipeline on Azure. Yet what makes it tick? And what makes it different from any of the cloud data warehouses

out there? In this section, we explore the primary use cases of Azure Synapse and how its architecture enables these patterns.

When to use Azure Synapse

Azure Synapse is optimized (down to the metal) for OLAP workloads over massive amounts of data. Similar to other cloud data warehouses, Azure Synapse really starts to shine in the multi-TB range, and only gets more efficient from there. Azure Synapse fits happily into any of these patterns:

- **Lift-and-shift of an on-premises data warehouse.**
- **Consolidating historical data from isolated silos across your company.**
- **Powering reports and dashboards across your company.**

WARNING

Do not try and use Azure Synapse (or any cloud data warehouse) as an OLTP database! Many of the performance improvements that make these data warehouse great at OLAP workloads, such as column store tables and distributed storage, also make them worse at the constant inserts, updates, and deletes. For OLTP workloads on Azure, Azure Synapse has a variety of configurations and options to check out.

OLAP VS OLTP

Online Analytical Processing (OLAP) and Online Transactional Processing (OLTP) help categorize the two major types of workloads in an enterprise data estate:

- OLTP represents short but frequent row-based transactions (INSERT, UPDATE, DELETE); operational data such as tracking the thousands of sales happening per second in an online marketplace.
- OLAP is defined by fewer but more complicated transactions, with complex operations such as aggregations, joins, and filters (column-based operations), used for business intelligence and predictive analytics.

Cloud-native architecture

Azure Synapse is comprised of a storage and compute layer, which are abstracted away from each other (this is what allows you to scale storage and compute independently of each other).

The storage layer is built over Azure BLOB Storage but uses an in-house columnar storage model for maximum compression and performance. As you load data into Azure Synapse, the system automatically distributes the data into 60 logical distributions; you can check this out yourself using the `dbcc_check_distributions` command. These distributions are further replicated, both locally and globally, to guarantee high availability even in the case of failover. When creating tables in Azure Synapse, you have the option to choose how a table is distributed — either round-robin (evenly and randomly), on a hash key (great for speeding up joins and aggregations on that key), or replicated across all 60 distributions (perfect for small dimension tables).

On the compute side, Azure Synapse consists of a control node, and a variable number of compute nodes based off your current scale factor, as shown in Figure 1-1. The control node handles connections and final aggregation of results, while the compute nodes are the elastic scale. The real intelligence is in the distributed query engine, which automatically breaks up the SQL query submitted to the control node and assigns tasks to the compute nodes. The distributed query engine in Azure Synapse automatically takes advantage of the performance optimizations you set up, including partitions, materialized views, and caching, meaning analysts and BI tools don't need to understand the complexities of cloud data warehousing to take advantage of the performance.


TECHNICAL STUFF

Any experienced DBA knows that a good data warehouse strikes a balance between building in performance optimization and giving controls to the user to optimize for their workload. While Azure Synapse handles the columnar storage, data distributions, query optimization, and caching automatically, it still gives an experienced DBA levers to fine-tune their workload, including:

- Statistics
- Table partitions
- Secondary indices

- Materialized views
- Ordered column store

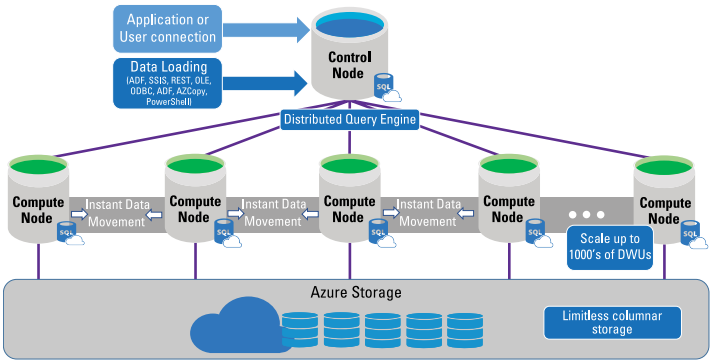

FIGURE 1-1: The compute side of Azure Synapse.

Considering the Must Haves of Any Data Warehouse

When selecting a cloud data warehouse, it's important to understand not only the strengths and weaknesses of each offering but also your enterprises' wants and needs. Based off your architecture and workload, some features may be deal breakers, while others you can live without. However, any enterprise data warehouse must address some common issues, such as security and resource management.

In this section, we walk through some of the most important areas to consider when choosing a data warehouse, including features and technologies to look out for.

Need for speed

Price performance has been one of the most important standards for data warehouses for decades, and that hasn't changed with cloud data warehouses. Business analysts need performant queries and interactive reports to be productive and answer critical business questions as quickly as possible. But on the other hand, every business has a budget, and only has so much to spend on a data warehouse. Therefore, the name of the game in data

warehouses is price performance — how much performance are you getting for your cost? While cloud data warehouses do enable a variety of price-perf skews through their elasticity, at the end of the day, once you settle on the scale factors and sizes that work for your company, it still comes down to raw price performance to maximize the value your analysts can get from your data.

Security, security, security

Security is always top of mind for any enterprise. A data breach irrevocably hurts customer trust and can result in fines, litigation, and lost business. When considering the security of your cloud data warehouse, you must consider security at every level, from the environment level (firewalls and integration with your identity management services such as Azure Active Directory) all the way down to granular data control (are features such as row and column level security built in or do they require manual workarounds?). Remember that the easier it is to secure your data warehouse, the less chance there is for gaps or mistakes in your security system, so look for built-in integration and security features that will make your data warehouse more secure (and keep your DBAs happy!).

Managing your workload

As mentioned at the start of this chapter, one of the greatest strengths of the cloud is the ability to easily throw more hardware at any given problem. But nothing in this world is free, and more scale always comes at more cost. Therefore, it's important to understand what tools your data warehouse gives you to not only control costs but also maximize your investment at a given scale factor. Consider the controls any potential DWaaS gives you to manage the differences in your workload: How do you deal with higher priority queries? What controls do you have over your concurrency? In general, how can you maximize your data warehouse usage without breaking the bank?

IN THIS CHAPTER

- Choosing data sources
- Differentiating the two types of data ingestion methods
- Keeping your warehouse data up to date
- Introducing Qlik Replicate (Formerly Attunity Replicate)

Chapter 2
Populating Cloud Data Warehouses

In this chapter we discuss the general approach and methods for populating cloud data warehouses with enterprise data.

Choosing Data Sources

As mentioned in Chapter 1, the enterprise data warehouse is a strategic repository that provides analytical information about the core operations of an enterprise. Consequently, deciding what data should live in the data warehouse can be a tricky but interesting challenge. Too few data sources can fail to produce insights, too many can be overwhelming. Answering the following questions helps to find the right ratio for your company:

- **Business stakeholders:** Who are the business stakeholders and what data do they need? How frequently do they need data?
- **Data model:** Is there an existing data model? Is it documented? Does it document all the business data? Are there gaps?

- **Database constraints:** Are there database constraints for your database table and objects that could impede data ingestion?

Drawing up a list of the business stakeholders and their data requirements helps you scope the effort and shortlist possible data ingestion methods. Your list can serve as an initial roadmap to identify and prioritize the most relevant source data systems. In addition, exploring any existing documentation enhances your understanding of data structure contents and limitations.

What are some likely source data systems? They could be:

- **Customer:** Your customer support system typically contains your master customer information. In addition, it can include any kind of new customer, or returning customer transactional information.
- **Finance:** Your finance system could include invoice data, fraud cases, or any other ERP generated financial information.
- **Marketing:** Marketing automation systems are the system of record for product campaigns, or special offers to your customers.
- **Inventory:** Warehouse and inventory management systems are the source for stock and supply chain information.
- **Sales:** What does your company sell? The CRM system is the master repository for customer information.

WARNING

Business data can reside in many different transactional enterprise applications including those on-premises and in the cloud. Additionally, different data fields in various systems may have the same definition, names, and data types. Therefore, it's critically important to cross check your requirements with your data model documentation and system metadata to ensure you're ingesting the correct data.

TIP

Using ingestion tools with metadata and lineage capabilities can help reduce confusion and improve the accuracy of downstream analytics reports.

Data Ingestion Methods

After you define what data is required, the source systems of record, the data model and frequency, you're ready to move onto investigating the best solution for data ingest. Typically, there are two categories: batch and real-time.

Batch ingest

Batch ingest has traditionally been the most common method of loading the data warehouse. It's generally a manual process that largely occurs on a schedule, pre-determined by the business stakeholder.

File import

Most cloud data warehouse projects begin by experimenting with different methods for importing a modest data set. Traditionally, these small data sets are formatted as comma separated (CSV) files and stored on corporate file shares. More recently, enterprises leverage the flexibility of cloud BLOB or file share for CSV storage. CSV import is the lowest common dominator ingestion method, which can be difficult to scale and operationalize.

Native loaders—bulk copy utilities

Many cloud data warehouses expand on the file import notion by providing native data loaders that can not only import files, but other simple data source types too. For example, many cloud data warehouses provide a bulk copy utility that can be used for migration or backup. However, while these utilities have great scalability, they're intended for ad-hoc use, rather than enterprise-production operationalization.

MIGRATING DATABASES TO AZURE

The Azure Database Migration Service is a fully managed service designed to enable seamless migrations from multiple database sources to Azure data platforms with minimal downtime. Visit `https://docs.microsoft.com/en-us/azure/dms/dms-overview?view=sql-server-2017` for more information.

Extract, Transform, Load

Extract Transform, Load (ETL) removes many of the downsides associated with file and native data loading tools by automating and operationalizing the entire data ingestion process. Data is extracted in batches from source systems on a regular schedule, transformed by business rules, and then loaded into a central data warehouse. ETL operationalizes the types of import performed by native loaders but requires dedicated technical programming resources.

Real-time data ingest

The second category of warehouse data ingestion is real-time data ingestion and is increasingly used by companies to improve responsiveness for business analytics. The concept involves loading transactions into the warehouse as and when they occur in the source systems. There are two categories of real-time ingest.

Micro-batch ETL

Micro-batch ETL is no different than traditional ETL, except that the batch schedule is set for minutes rather than hourly. A micro-batch process queries the source system every few minutes to determine what new data to insert into the warehouse.

Change data capture

Change data capture (CDC) technology identifies and captures just the most recent production data and metadata changes that the source has registered during a given time period, typically measured in seconds, then automatically replicates a copy of those changes to the data warehouse.

REMEMBER

Like micro-batch, CDC copies data from one source to one target, or one source to multiple targets. However, it doesn't incur the overhead of additional database queries.

In addition, CDC also identifies and replicates any changes to a data source schema (DDL changes), enabling data targets to dynamically adapt to structural updates without manual intervention.

CDC follows the Extract, Load, Transform (ELT) pattern where data is extracted from the source systems as and when the change occurs, then immediately loads the target data warehouse without any transformation. Transformation occurs later, usually when the data is consumed by an ensuing analytics process.

ETL VERSUS ELT

Extract, Transform, Load (ETL) is the traditional method of moving data from a source system into a data warehouse. The process involves extracting data from outside sources in batches, transforming it to fit operational needs (sometimes using staging tables), then loading it into the end target data warehouse. This approach is reasonable so long as data volumes are modest and can be handled within the time allotted for the batch process.

Extract, Load, Transform (ELT) is a modern alternative to ETL. ELT extracts raw data from the source and loads the target data warehouse, typically in staging tables. The data is transformed in the future by other processes. ELT has the advantage in that it's extremely high performance, because no additional transformation overhead occurs and is ideal for handling massive data volumes.

CDC has fundamental advantages over batch ETL, namely:

- **Performance:** CDC is ideally suited for high-frequency transaction systems because no intermediary data transformation occurs. Data is quickly replicated from source to target.
- **Data accuracy:** Data in the warehouse is always accurate and up to date because source data is copied to the target as soon as a change occurs.
- **Low overhead:** Log-based CDC doesn't add additional processing overhead to source systems and minimizes disruption to production workloads.

Together, these advantages enable IT organizations to meet the real-time deliverability, efficiency, and scalability requirements of a modern data architecture.

Keeping Warehouse Data Current

For decades, data warehouses such as IBM Netezza, Oracle Exadata, and Teradata have served as the system of record for reporting and analytics. However, the emergence of new cloud-based alternatives such as Microsoft Azure Synapse has highlighted the importance of

analyzing data in real time. As a result, CDC has become a critical building block of modern data warehouse architectures.

CDC uses several methods to instantly deliver replicated data to the data warehouse target.

- **Transactional delivery:** Data updates are copied in the same sequence in which they were applied to the source. This method is appropriate when sequential integrity is more important than ultra-high performance. For example, daily financial reports need to reflect all completed transactions at a specific point in time, so transactional CDC would be appropriate.
- **Aggregated delivery:** Multiple updates are bundled together and sent to the target system(s). This facilitates the processing of high volumes of transactions when performance is more important than sequential integrity. This method is useful for aggregate trend analysis based on the most data points possible. Aggregated delivery also integrates with certain target data warehouses' native utilities to apply updates.
- **Stream-optimized delivery:** This method replicates source updates into a message stream that is managed by platforms such as Apache Kafka or Azure Event Hubs, before landing into the data warehouse. Stream-optimized delivery is important for new initiatives such as machine learning and Internet of Things (IoT) that require data in motion rather than at rest.

CHANGE DETECTION MECHANISMS

There are three approaches to change detection, with some being significantly more beneficial than others: triggers, queries, and log readers. Triggers and query-based CDC can impose additional overhead because they require the database to perform extra work. Log-based CDC, however, doesn't suffer this penalty because it works by scanning for changes in the database transaction log files that already exist for backup and recovery purposes.

Leveraging these CDC delivery methods ensures that the data in your data warehouse are always accurate and up to date.

Qlik Replicate Overview

Qlik Replicate (formerly Attunity Replicate) is a modern data integration platform built on CDC that provides a highly automated and consistent mechanism for replicating incremental data updates while minimizing production workload impact. Qlik Replicate integrates with all major databases, data lakes, data warehouses, streaming, cloud, and mainframe systems.

With Qlik Replicate, you can address use cases that include the following:

- Real-time data warehouse ingestion and update
- On-premise database to cloud database migration
- Database transaction publication to message streams
- Mainframe data offload and modernization

REMEMBER

Qlik Replicate's CDC mechanism captures data and metadata changes through the least-disruptive method possible for each specific source, which in most cases is a log reader.

Changes are sent in-memory to the target, with the ability to filter rows or columns that don't align with the target data warehouse schema or user-defined parameters. Qlik Replicate also can rename target tables or columns, change data types, or automatically perform other basic transformations that are necessary for transfer between heterogeneous end points, as shown in Figure 2-1.

Qlik Replicate capabilities include the following:

- **Automation:** A web-based console enables data architects rather than ETL programmers to configure, control, and monitor replication tasks, thanks to a self-service approach that eliminates ETL scripting. The 100 percent automated process includes initial batch loads, transitions to CDC, target schema creation and schema/DDL change propagation,

improving analytics agility, and eliminating the need for cumbersome recoding of brittle manual processes.

- **Efficient and secure cloud data transfer:** Qlik Replicate can compress, encrypt, and transfer data in parallel streams into, across, and out of cloud architectures, including the Microsoft Azure Data Platform. Qlik technology has powered over 85,000 data migrations to the cloud and is recognized as the preferred solution by Microsoft.
- **Rapid ingestion for cloud data warehouses:** Qlik Replicate provides high performance data loading to cloud data warehouse targets such as Azure Synapse through native CDC APIs.
- **Universal data connectivity:** Qlik Replicate supports over 75 data sources and targets including relational databases, data warehouses, data lakes, enterprise applications, and streaming systems, in on-premises, cloud, or hybrid architectures.

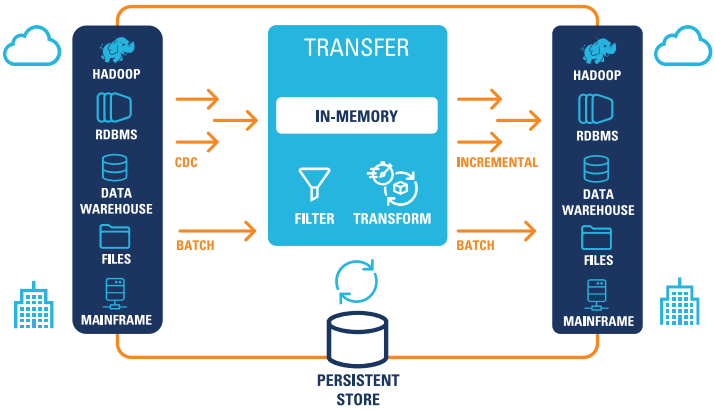

FIGURE 2-1: Qlik Replicate.

In addition, the Qlik Enterprise Manager™ (formerly Attunity Enterprise Manager) allows you to

- Design, execute, and monitor thousands of Qlik Replicate tasks across highly distributed environments.
- Monitor alerts and analyze system KPIs in real time.

- Visualize data flow trends to assist with capacity planning and performance monitoring.
- Provide REST and .NET APIs to integrate with workflow automation systems and applications built on microservices architectures.

Qlik's software enables continuous, efficient, and low-impact data integration that automates key aspects of data warehouse ingestion for analytics. It helps you maintain consistent, flexible control of data flows throughout your enterprise while reducing dependence on highly skilled and costly ETL programmers.

IN THIS CHAPTER

- Taking common data warehouse problems into consideration
- Defining agile automation
- Exploring the data pipeline principle
- Adopting agile methods
- Introducing Qlik Compose (formerly Attunity Compose) for Data Warehouses

Chapter 3
Agile Data Warehouse Automation

Agile data warehouse automation is more than just automatically producing the necessary scripts to build a data warehouse. In this chapter, we explain why you should explore agile data warehousing to automate the entire data warehousing life cycle.

Considering Common Data Problems

Azure Synapse offers enterprises a compelling set of benefits including high performance analytic queries at lower total costs of operation than traditional on-premises data warehouses. However, data warehousing projects can still fall short of expectations despite the obvious flexibility and value that Azure Synapse provides. This is especially true if warehousing projects can't adapt to keep pace with rapidly changing analytics and business requirements.

However, it's unfair to blame the data warehouse technology. For years, the most significant failing of the data warehouse philosophy was the necessity to design, build, and deploy data models

that could satisfy every conceivable reporting eventuality! Get the up-front data model right and reporting is easy. Get the data model wrong, with missing fields or wrong relationships, and insights are impossible to derive. Furthermore, rearchitecting a data warehouse to address these shortcomings is a tedious, technical, and laborious process.

Similarly, business user demands for increasing volumes of data supplied by an ever-expanding list of back-office data sources only serves to amplify delivery problems. As a result, frustrated business teams pursue self-service options that encourage the growth of isolated data hubs. These "quick fixes" create data inconsistencies, and generally suffer from a lack of governance and data oversight.

To complicate matters, data architects, database and data warehouse administrators, and integration engineers often used a differing variety of incompatible database modeling and integration tools to collect, prepare, and distribute the right information for consumption by business analysts and executives. It impeded collaboration and increased inefficiency that ultimately slowed down data warehouse project delivery.

Also, IT typically follows a traditional waterfall approach to designing, building, and operating the data warehouses and many practitioners continue this trend as they move to the cloud. It's no wonder that data warehouse projects appear to move at a glacial pace as they roll into production.

Fortunately, there is an alternative to help regain control, improve responsiveness, and increase efficiency. Data warehouse automation — an underutilized methodology — is a proven and effective way to resolve the data delivery challenges of data warehousing.

Introducing Agile Automation

Data warehouse automation is more than just deploying technology for efficiency's sake, and more than just simply automating ETL scripts. It's wholistic approach automates the entire data warehousing life cycle from planning, analysis, and design through to operations, maintenance, and change management.

A modern agile data warehouse approach relies on three concepts:

- It provides a central mechanism and framework to organize the diverse tasks that are critical to the process of designing, developing, and deploying a data warehouse.
- Data warehouse automation provides an abstraction layer that masks the complexities of warehouse data design, development, and maintenance.
- Data warehouse automation is an approach to manage the lifecycle of your data warehouse systems. It addresses tasks including
  - Maintaining an existing data warehouse while containing costs.
  - Upgrading a warehouse while ensuring continuity.
  - Migrating a warehouse from one to another across clouds if necessary.
  - Eventually retiring a data warehouse when the time is right.

Adoption of data warehouse automation should change the mindset of managing data warehouses. The widely accepted practice of extensive up-front analysis, design, and modeling moves from "get it right the first time" to "develop fast and develop often."


TIP

Although the phrase "agile data warehouse automation" is similar to "agile development," it doesn't mean that you must adopt an accepted agile process. You can achieve substantial speed and quality without fully embracing and implementing a formal agile methodology.

Agile automation benefits

Agile data warehouse automation has many of the same benefits that you see in other verticals or disciplines, such as manufacturing, when you embrace it:

- **Reduction of manual effort:** With less time spent on manual tasks, new data warehouses can be quickly deployed. This is especially important for cloud data warehousing where instances are dynamically created, populated, and destroyed for specific analytics uses.

- **Increased productivity:** Adding new data sources and modifying models for existing, persistent data warehouses takes virtually no time at all, ensuring a faster time to insight.
- **Improved data quality and consistency:** Automation with no manual intervention reduces the possibility of human error and consequently improves data quality. In addition, life cycle and artifact metadata can be automatically managed to provide complete data lineage.

With an agile data warehouse automation, you can now match the level of agility and flexibility typically associated with modern cloud computing.

KIMBALL VERSUS INMON VERSUS LINSTEDT

Data warehouse design has evolved different approaches though the years.

Ralph Kimball's approach stressed the importance of data marts, which are repositories of data belonging to lines of business. The data warehouse is simply a combination of different data marts that facilitates reporting and analysis. The Kimball data warehouse design uses a "bottom-up" approach.

Bill Inmon regarded the data warehouse as the centralized repository for all enterprise data. In this approach, an organization first creates a normalized data warehouse model. Dimensional data marts are then created based on the warehouse model. This is known as a top-down approach to data warehousing.

Dan Linstedt offered yet another alternative. Linstedt's data vault method looks at historical data and focuses on issues such as auditing, loading speed, and resilience to change as well as emphasizing the need to trace the data origin. Therefore, every row in a data vault must be accompanied by record source and load date attributes to enable auditing that can trace values back to the source.

Agile data warehousing is not a design philosophy but a methodology, and consequently can be used with any of the former design approaches.

Data Pipeline Principles

Data modelling, design patterns, metadata, and reuse are all cornerstones of agile data warehouse automation. However, one key concept is crucial to success. Establishing an automated data pipeline that can stream data from source to analytics-ready data marts is essential.

Without pipelining

Many of the data warehouse problems mentioned earlier in the chapter occur regardless of design philosophy and are largely due to the traditional data flow shown in Figure 3-1.

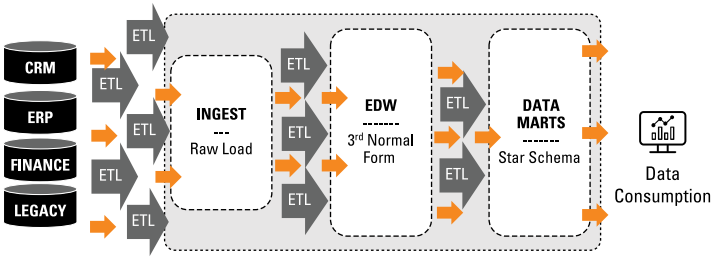

FIGURE 3-1: A traditional data flow.

Raw data flows from source systems on the left to eventual consumption on the right, aided by hundreds or thousands of different ETL process in between. Often each ETL process is isolated, serves a singular purpose, and requires dedicated management. The resulting architecture is extremely brittle, complicated, fragile, time consuming, and error-prone.

It's not surprising that data warehouses that adopt this architecture take years or even months to implement, are difficult to adapt to new requirements, and are extremely challenging to manage.

DATA WAREHOUSE AUTOMATION VERSUS TRADITIONAL ETL

A common question that arises is whether data warehouse automation (DWA) is just another name for a specialized extract-transform-load (ETL) process.

DWA is a singular approach optimized for automatically building and managing the code associated with the data warehouse lifecycle. For example, DWA automatically builds transactional, periodic snapshots and accumulates snapshot fact tables; Type 0, 1, and 2 slowly changes dimensions; and automatically moves data from multiple data sources through various architectural zones into a consolidated data mart consumption zone.

ETL, on the other hand, is a general-purpose programming tool for moving and transforming data between different systems. It provides components, in one form or another, which you manually assemble to accomplish one or more tasks. You could try to automate slowly changing dimension fact tables with an ETL tool, but it would be like writing complex calculus in a spreadsheet — time consuming, not well understood, and prone to error.

Establishing an agile data pipeline

Many data practitioners feel that a new approach is required to establish a data pipeline rather than retrofitting legacy ETL. For a data pipeline to be more robust it must support:

- **Leveraging a multi-zone architecture:** A pipeline should leverage the best practice of using different logical zones for various stages of the data flow.
- **Embracing data modelling:** A core component of data warehouse design, but tooling should be integrated and not an isolated component.
- **Integrating data ingestion:** High performance data loading and updates should be an integral feature and not thousands of separate tasks.

- **Generating scripts automatically:** Transformation scripts for different types of denormalized data structures (star schemas, OLAP cubes) used for BI reporting should be automatically generated from data models.
- **Documenting dynamically:** Ideally schemas, scripts, and data attributes should be automatically generated and self-documented from models. Documentation should dynamically adapt as logical data models change.
- **Provisioning self-service data mart:** The pipeline should support self-service data mart provisioning for BI analysts.

In an ideal world, an agile data pipeline looks like Figure 3-2. It leverages the best practice of using different logical zones for the data warehouse architecture and automates the various functions in between.

FIGURE 3-2: An agile data pipeline.

For example:

- **Landing:** The landing zone contains just the raw data updates from the source systems.
- **Staging:** The staging zone is used to collate, standardize, enrich, and prepare data for loading into the data warehouse.
- **Warehouse:** Data in the warehouse zone is typically transformed to traditional third normal form.
- **Marts:** Data marts are specific data sets used for reporting purposes and often conform to certain schema designs such as star or snowflake.

WHAT IS A STAR SCHEMA?

Another common question is "What is a star schema and why is it important for data warehouse projects?" A star schema is most widely used to develop data warehouses and dimensional data marts. It consists of one or more "fact tables" referencing any number of "dimension tables." The star schema gets its name from the model's resemblance to a star shape with a fact table at the center and the dimension tables surrounding it that resemble the star's points. They're denormalized and simplify common business reporting logic.

Knowing When to Adopt Agile Automation

It's understandable that there is a reluctance to adopt a new methodology because you have so much invested in traditional tools and methods. Your teams are already trained, your warehouse is already built, and your operational processes are all in place. However, you also know that you can't keep pace. The eternal question is therefore, "When is the best time to adopt agile data warehouse principles?"

Natural inflection points

There are several inflection points that are a natural time to adopt agile automation; namely:

- **Cloud migration initiative:** Many enterprises looking to increase flexibility, manage costs, and redirect resources are moving their data warehouse to the cloud. This is a good opportunity to review your technology base and leverage a more modern and responsive approach.
- **Increasing data delivery times:** When business analysts and BI teams can't get the data they need in a timely manner, you should consider moving to a more responsive approach.

- **End of life:** It's a truism that nothing lasts forever, including foundational infrastructure technologies. When a crucial warehouse technology becomes obsolete, it's a perfect opportunity to consider a more modern alternative for accelerated data delivery.
- **Economic drivers:** Sometimes a vendor technology change forces a move. For example, the change to S4HANA is pushing many enterprises to reevaluate the economics of HANA BW and look for a lower cost data analytics platform.
- **Resource reduction:** Data warehouses built using traditional methods and tools are resource-intensive and expensive. When a data warehouse team becomes budget or resource constrained, look to automation to relieve the resource pressures.
- **Adopting new innovations:** New projects oriented around the data warehouse meet new business requirements. Rather than continue to use the same approach and experience the same problems with traditional data warehousing, companies with new requirements or who have time-sensitive business requirements to meet can benefit from this new, more automated approach.

Qlik Compose for Data Warehouses

Qlik Compose (formerly Attunity Compose) is an automation solution for data architects who design, deploy, and operate agile data warehouses. With Qlik Compose you can accelerate the business time-to-value for your data warehouse projects by reducing the time and the resources necessary to implement a data warehouse, data mart, or data hub.

Qlik Compose virtually eliminates the manually coded, time-consuming, and error prone repetitive tasks of data ingest and modelling while providing the benefits of an accelerated data delivery architecture to both on-premises and cloud environments, as shown in Figure 3-3.

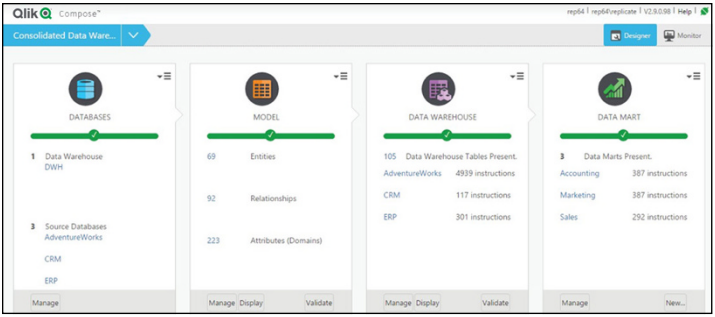

FIGURE 3-3: Qlik Compose.

Qlik Compose addresses the following use cases:

- Replacing a brittle legacy warehouse or data mart
- Migrating and re-platforming a legacy warehouse to the cloud
- Augmenting an existing data solution with a new data warehouse
- Consolidating data sources and new analytics initiatives
- Prototyping new analytics environments

Qlik Compose uses key innovations and technologies to enable efficient, effective data warehouse design and lifecycle management. It has two categories of data warehouse automation features:

- Design-time focuses on improving the productivity of creating and deploying a data warehouse.
- Run-time features focus on the efficient operation of the data warehouse after it's deployed.

Design-time automation

Qlik Compose provides the following automation features to eliminate the error-prone manual coding associated with the repetitive steps of data warehouse design and implementation.

Automation features include:

- **Model-driven or data-driven:** The visual modeler designer allows you to generate data warehouse models in three ways:

- You could use the native graphical tool to define entity relationship diagrams from scratch.
- You could import models generated by third-party tools such as CA Erwin.
- You could discover models by reverse engineering the source databases. You can leverage any or all the different data warehouse design styles including Inmon, Kimball, and Data Vault.

- **Real-time source data ingestion and CDC updates:** Through integration with Qlik Replicate, Qlik Compose for Data Warehouses offers support for the industry's widest ecosystem of database sources, with real-time and query-based change data capture (CDC).
- **Physical data warehouse management:** After source metadata is discovered and mapped, and the model generated, then the staging, data warehouse, and data mart structures are generated with a click of a button.
- **Data mart generation:** Qlik Compose supports three types of data marts out of the box: *transactional* (a star schema with a transactional fact table allows you to create "history" type reports, as shown in Figure 3-4), *aggregated* (a star schema to support "summation" reports), and *state-oriented* (a star schema that supports "period" reporting between times).

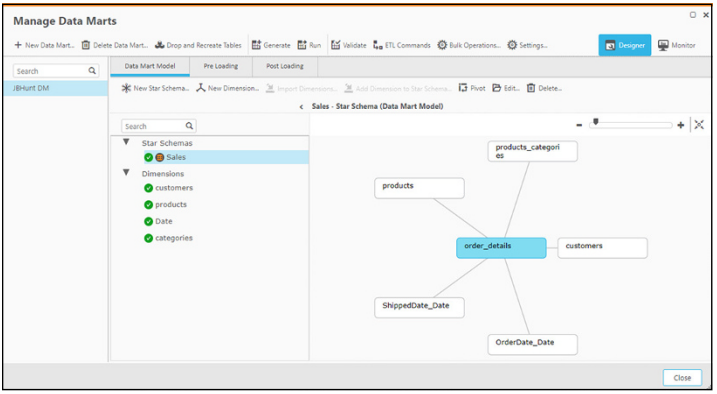

FIGURE 3-4: A star schema in Qlik Compose.

Run-time automation

Qlik Compose also provides the following features for ongoing operation and maintenance of the data warehouse:

- **Workflow designer and scheduler:** A built-in workflow designer is provided to enable you to build and schedule your data warehouse and data mart ETL tasks to run as a single, end-to-end process.
- **Monitoring:** The monitor view shows the current status of workflows and tasks that are active and running.
- **Notification:** Users can be notified about various operational events that occur as warehouse tasks are running.
- **Data profiling and quality enforcement:** Data profiling features validate data to preemptively identify format discrepancies such as null or blank fields. In addition, data quality rules can be configured to enforce exception policy before loading data.
- **Lineage and impact analysis:** System metadata is automatically generated and provides a visual representation of the data flow pipeline. In addition, integrated impact analysis can identify how schema changes can impact the data warehouse.
- **Automated documentation:** Metadata is leveraged to automatically create documentation that self-describes the data warehouse environment.
- **Lifecycle management:** Automated development, testing, acceptance, and production (DTAP) capabilities streamline migration of data warehouse projects between different environments. Deployment packages that contain models, ETL scripts, and other components are generated by a single button click.

REMEMBER

Agile data warehouse automation is more than just automatically producing the necessary scripts to build a data warehouse. You should implement solutions and processes that encompass the entire data warehousing life cycle.

IN THIS CHAPTER

- Ferguson Enterprises
- Save the Children

Chapter 4
Case Studies

In this chapter, we show how two customers are leveraging solutions from Microsoft and Qlik for their Azure Data Pipelines.

Ferguson Enterprises

Ferguson Enterprises is raising the bar for industry standards as the top-rated wholesale supplier of commercial and residential plumbing supplies. Their expertise goes beyond plumbing. They're a diverse distributor that spans multiple businesses, including HVAC/R, waterworks, and industrial. In the past 65 years, they've grown from a local distributor to a $16.7 billion dollar company with more than 1,400 locations and over 27,000 associates nationwide. Their associates provide expert advice and a range of products and services customers want to improve their construction, renovation, and maintenance projects. Providing world-class customer service is a cultural belief that is demonstrated every day through their expansive product selection, backed by knowledgeable associates.

Challenge

Ferguson wanted to rearchitect their siloed, on-premise data environment that included Netezza, Oracle Exadata, and other sources so that business units could share data more easily and make more insightful decisions in the moment. But getting an analytics project delivered took months, sometimes even years, to complete. The technologies in play weren't scalable, and it was a challenge to find resources that could leverage them effectively.

As Brandon Lassiter, Senior Director of Enterprise Information Management at Ferguson said, "When I came into the organization, our stakeholders were pretty dissatisfied with how things were going. Simple projects took months because of complicated architecture and siloed resources. To compound the issue, the platform was down at least once a week. On my very first day, we had a major incident. In essence, we spent a lot of money on solutions that just didn't meet our needs."

Solution

They chose the Azure Data Platform and Qlik software as their new solution. Qlik Replicate was able to migrate data from multiple sources in real time to the Azure Data Lake Store for multiple enterprise data sources. Additionally, Qlik Compose was able to automate the ETL processing previously built into PL/SQL and DataStage. Lastly, Qlik Enterprise Manager is providing a dashboard for the process.

Brandon Lassiter said, "Qlik (Attunity) Replicate, in particular, is a huge game changer in three ways: 1) It has low overhead on our infrastructure. We see very little latency in our databases, and it doesn't require installing agents on endpoints; 2) The time from install to value is *fast*. We started a proof of concept (POC) and were replicating data to our Azure environment in about an hour. The team ramp up was very quick; and 3) Qlik (Attunity) is invested in our success. When we were making the decision to move all of our analytics capabilities to Azure, I knew I needed not just a vendor, but a partner that would listen to feedback, and adjust to our needs. Qlik (Attunity) met or even exceeded my expectations almost every time."

Brandon Lassiter sums it up like this, "All good companies understand the value of unlocking the potential of their data. The problem is, we've all read the articles on many analytics projects

being a total failure, due to complexity, cost, etc. For me, the value proposition of Qlik (Attunity) is to simplify the first step, getting the data out. But they've taken it a step further. With Qlik (Attunity) Compose, we can now transform the data and deliver enterprise models for more traditional data warehouse needs. Because the software is intuitive and easy to implement in our Azure environment, the time to value is a fraction of our old on-premise solutions."

Save the Children

Save the Children believes every child deserves a future. Since their founding 100 years ago, they've changed the lives of more than 1 billion children. In the United States and around the world, they give children a healthy start in life, the opportunity to learn, and protection from harm. They do whatever it takes for children — every day and in times of crisis — transforming their lives and the future they share. More information can be found at `http://www.savethechildren.org/`.

Challenge

The IT Business Solutions Group located in Fairfield, CT was working on a new business intelligence (BI) initiative. Santha Kumar, Director of Application Architecture and Development, was leading the project to move data onto a data lake in the cloud so that the data analysts could use it to measure program and financial data across the Agency. To do this, the IT team at Save the Children needed to work with real-time data from its CRM and ERP systems without any downtime, rather than a snapshot of yesterday's data.

While Save the Children primarily uses Microsoft technology, the IT Business Solutions Group had tried using other tools for some of its ETL work. Unfortunately, it wasn't proving to be the right solution for moving bulk data.

Solution

They selected Qlik Replicate as the data integration software to move its data from multiple sources to Azure Data Lake Storage (ADLS) and Azure Synapse. Now, the data analysts at Save the Children will have insight into their data in real time. Data are

no longer siloed. They also can run up-to-the-minute reports to help them measure various areas of their business and react quickly to address issues. Save the Children is also looking at Qlik Compose to automate their data warehouse.

"With Qlik (Attunity) we're able to better integrate data multiple sources — including our CRM and ERP systems. Working tightly with Microsoft technology, Qlik's (Attunity's) solutions make our data available immediately so that data analysts across Save the Children have access to the freshest data possible for real-time analytics and reporting." — Santha Kumar, Director of Application Architecture and Development, Save the Children.

Chapter 5
Ten Resources

Here we present you with ten resources that provide more information to help you learn more about data warehouse automation for Azure Synapse Analytics.

- Azure Synapse Analytics Overview and Walkthrough with Microsoft Mechanics `https://www.youtube.com/watch?v=tMYOi5E14eU&t=2s`
- Azure Synapse Analytics Homepage `https://azure.microsoft.com/en-us/services/synapse-analytics/`
- Azure Synapse Analytics Free Trial: `https://azure.microsoft.com/en-us/free/synapse-analytics/` get started with 12 months of free service to accelerate and optimize your analytics with a fully-managed data warehouse
- Sign up for the Azure Synapse Analytics preview `https://azure.microsoft.com/en-us/services/synapse-analytics/request-for-preview/`
- Azure Synapse Tips and Best Practices `https://docs.microsoft.com/en-us/azure/sql-data-warehouse/cheat-sheet`
- Video: Save the Children Moves Real-time Data to Azure Data Lake with Qlik: `https://www.youtube.com/watch?v=8tvt9fw7_p0`

- Case Study: BITMARCK Uses Qlik and Microsoft to Streamline Data Integration and Enable Self-Service, Real-time Business Intelligence for Customers. https://www.qlik.com/us/-/media/files/resource-library/global-us/direct/case-studies/cs-qlik-replicate-for-bitmarck-case-study-en.pdf
- Mainframe Data Integration with Qlik Replicate and Microsoft Azure https://www.qlik.com/us/-/media/files/resource-library/global-us/direct/datasheets/ds-qlik-mainframe-integration-for-azure-en.pdf
- Easy Data Ingestion and Automation: Qlik Data Integration for Micosoft Azure https://www.qlik.com/us/-/media/files/resource-library/global-us/direct/datasheets/ds-qlik-solutions-for-azure-sql-data-warehouse-en.pdf
- Microsoft Azure – Database Migration and Data Streaming https://www.qlik.com/us/products/technology/microsoft

Stream real-time data to Azure Synapse

As you make the decision to move your data warehouse from on-premise to the cloud or cloud to cloud, there are many things to take into consideration. You need to take into account the differences that exist between an on-premise data warehouse and a cloud data warehouse. In this book, you can find out how a cloud data warehouse in Azure has advantages in cost, time to value, and the ability to work with real-time data across the organization for analytics.

Inside...

- Learn what you need to build your enterprise data warehouse in Azure
- Move the right data at the right time
- Sidestep common integration problems
- Take an agile approach to data warehousing

Qlik LEAD WITH DATA | Microsoft

Go to Dummies.com®
for videos, step-by-step photos, how-to articles, or to shop!

ISBN: 978-1-119-62091-4
Not For Resale

for dummies® A Wiley Brand

Also available as an e-book

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.