Brought to you by:

TRIFACTA®

Data Preparation

for dummies®
A Wiley Brand

Find out what data preparation is

Compare data prep to other solutions

Implement data prep at your company

Trifacta Special Edition

Ulrika Jägare

About Trifacta

Trifacta is the global leader in data preparation. Trifacta leverages decades of innovative research in human-computer interaction, scalable data management, and machine learning to make the process of preparing data faster and more intuitive. Around the globe, tens of thousands of users at more than 10,000 companies, including leading brands like Deutsche Boerse, Google, Kaiser Permanente, New York Life, and PepsiCo, are unlocking the potential of their data with Trifacta's market-leading data preparation solutions. Learn more at **trifacta.com** or **www.trifacta.com/data-wrangling**.

Data Preparation

Trifacta Special Edition

by Ulrika Jägare

for dummies®
A Wiley Brand

Data Preparation For Dummies®, Trifacta Special Edition

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

For general information on our other products and services, or how to create a custom *For Dummies* book for your business or organization, please contact our Business Development Department in the U.S. at 877-409-4177, contact info@dummies.biz, or visit www.wiley.com/go/custompub. For information about licensing the *For Dummies* brand for products or services, contact BrandedRights&Licenses@Wiley.com.

ISBN: 978-1-119-70156-9 (pbk); ISBN: 978-1-119-70158-3 (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: Carrie Burchfield-Leighton
Sr. Managing Editor: Rev Mengle
Acquisitions Editor: Katie Mohr
Production Editor: Tamilmani Varadharaj
Business Development Representative: Karen Hattan

Table of Contents

Introduction

Today's data is more diverse and complex than ever before. It's time-consuming and technically challenging to prepare the data into a format suitable for analysis. The awareness that data is not only important but also in fact your company's most valuable asset is growing fast in the industry. Data is vital for ensuring that organizational information is accurate, timely, complete, cost-effective, and accessible, and that it enables you to take proactive and conscious decisions throughout the business.

Data is the foundation of business information and knowledge and ultimately the wisdom for correct decisions and actions. If the data is relevant, accurate, meaningful, and actionable, it helps in the growth of the organization. If not, it can prove to be useless and even harmful to a scaling enterprise.

Therefore, treating your data correctly becomes a fundamentally important task. Getting your data preparation, or *data wrangling* as it's also referred to in this book, right is essential in order to increase the quality of the data and information. Ultimately, it's all about making it possible for you and your company to be able to effectively use and rely on the data at hand.

About This Book

This short book is packed with useful information about data preparation. In this book, you not only learn about the shift from desktop or on-premises data preparation solutions to cloud-based platforms but also what the main principles of data preparation are all about.

You discover that data preparation is no longer a task just for accomplished data engineers or one that requires coding skills, and I give you a bit about the new cloud-based solution from Trifacta and how it democratizes data preparation, making it achievable for anyone.

Icons Used in This Book

I occasionally use special icons to focus attention on important items. Here's what you find:

REMEMBER

This icon reminds you about information that's worth recalling.

TIP

Expect to find something useful or helpful by way of suggestions, advice, or observations here that help you leverage experiences from other implementations.

WARNING

Warning icons are meant to get your attention to steer you clear of potholes, money pits, and other hazards. Paying extra attention to these parts in the book help you avoid unnecessary roadblocks.

TECHNICAL STUFF

This icon may be taken in one of two ways: Techies will zero in on the juicy and significant details that follow; others will happily skip ahead to the next paragraph.

Beyond the Book

This book can help you explore general strategies for how to approach data preparation in your company. However, this book is a relatively short introduction to data preparation, so for further reading and deep dives on the topic, the following books and articles are recommended:

- **Data preparation vs. ETL in the cloud:** `www.trifacta.com/gated-form/eol-etl-data-wrangling-future-cloud`
- **Self-service data preparation for messy files:** `www.trifacta.com/gated-form/integrating-unfamiliar-data-ebook`
- **Learn more about Trifacta:** `www.trifacta.com`

IN THIS CHAPTER

- Explaining data management using traditional ETL
- Using Excel/manual coding in data preparation
- Sorting out how SQL/in-database coding works
- Listing benefits and drawbacks
- Using data preparation functionality
- Exploring flexibility in cloud-based solutions

Chapter 1
Exploring Different Approaches to Data Preparation

There is a common misperception that data analysis is mostly a process of running statistical algorithms on high-performance data engines. In practice, this is just the final step of a longer and more complex process where 80 percent of an analyst's time is spent wrangling data to get it to the point at which this kind of analysis is possible. Not only does data wrangling consume most of an analyst's workday, but also it represents much of the analyst's professional process. It captures activities like understanding what data is available, choosing what data to use and at what level of detail, understanding how to meaningfully combine multiple sources of data, and deciding how to distill the results to a size and shape that can drive downstream analysis.

REMEMBER

Many companies have invested a lot of time and money into the notion of putting all their data in one common storage location and then thinking all their problems will be solved. However, companies soon discover that despite all their efforts, the data is still difficult to find, access, and use. Succeeding with data management for analytics, reporting, and machine learning is clearly about a lot more than just data storage.

You can perform data preparation through many methods, tools, and techniques that range from being manual to highly automated and efficient. This chapter aims to describe different ways to perform data preparation, including benefits and limitations with each approach.

Describing How Legacy ETL Works

Extract, Transform, Load (ETL) is a commonly used term across the industry. It refers to the process of

- **Extraction:** Extraction is pulling data from many sources (traditionally relational databases). The data collection can be done as full extraction or partial extraction.
- **Transformation:** Data is transformed to ensure consistency in analysis. This process typically includes changing the data's format; standardizing values such as currencies, units of measurement, and time zones; enriching and validating the data to eliminate missing values and duplicates; and applying business rules.
- **Loading:** This includes loading and writing the data into the targeted storage unit: the database for use in an application, business intelligence solution, or data analysis product.

An overview of the traditional ETL process is described in Figure 1-1. The process is linear and usually assumes IT responsibility for ETL activities in an organization. Legacy ETL is slow and requires many iterations.

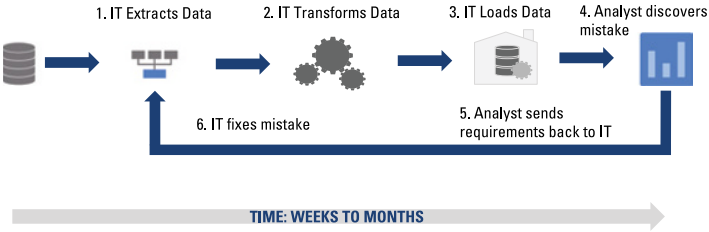

FIGURE 1-1: A traditional ETL process.

For a relatively long time, businesses have been relying on IT to run the ETL process to get a consolidated view of their data, not by the individuals who understood the data best. To ensure that the data is properly prepared and can be relied on for making better business decisions, it's vital to understand the data and the business context. Getting ETL right is still a core component of an organization's data integration system.

REMEMBER

For many years, traditional ETL was the only way to get data ready for analysis. The ETL process, however, comes with its own challenges and flaws that can potentially contribute to various sets of losses in any ETL activity.

Integrating data across different sources is challenging. It entails programming of scripts to parse the source data. If standard drivers aren't available, coding will be needed to complete the desired function.

WARNING

Building a representative architecture for an ETL project can also be tricky because you can't actually see the data in ETL processes. Going straight to coding without taking into consideration the overall bigger picture can cause serious problems for your team performing an ETL job.

The quality of data and various types of dependencies that exist in the data can impact the ETL process, as well as the complexity of the data relationships. When accessing data from different systems and moving data into the cloud, the quality of the data can't always be ensured. The data may be inconsistent, too, generating even more delays and cost to the ETL activity. However, once the data moves to the cloud environment, data preparation can be used to transform it for further use.

Compatibility of the data source and target data (to merge with) and scalability of the ETL process are other common technical challenges. Scalability can be a tricky issue that you may come across, and it depends on the size of the data you're dealing with. There can be operational changes in source data systems and ongoing revisions to target schema definitions and scope. This all adds complexity to the ETL process, although scalability limitations can usually be addressed through a cloud-based architectural setup.

Sorting Out Excel/Manual Coding

Although Excel isn't the optimal way of doing data preparation, it is still a widely used tool. This is especially true when the dataset is small and when the purpose is more of a one-time data preparation exercise. Once you need to scale-up your datasets and expand your data preparation activities to be spread over several teams, you quickly realize that manual coding Excel does not really scale, or support team collaboration efforts. The truth is that collaboration in Excel is basically impossible.

Since the only way to get value out of your data is to first prepare the data properly, it's important to know that the most time-consuming part in Excel is data cleansing using manual coding. It's usually extremely slow and difficult. However, this step is important because the cost of a mistake caused by incomplete information, discrepancies, and outliers can cause serious faults in your analysis that could significantly impact business outcomes. Remember that keeping track of *data lineage,* meaning your data origin, what happens to it, and where it moves, is a vital part of data preparation.

WARNING

Unfortunately, there are often unrealistic expectations on how long data preparation should take. Your manager may think that you can click a few buttons to transform a raw dataset into actionable analysis within an hour or two, but the reality is, no matter how powerful Excel is, it can still take several hours and more to manually compile and clean your data using spreadsheets. And if more complex coding or programming is needed to complete your analysis, you may have to take an online tutorial to learn how to perform a task or involve the IT department, both of which can add time and effort that further add to the time it takes to get the data cleaned.

WARNING

During the analysis process, spreadsheets continuously evolve, increase in complexity, and become more susceptible to errors. Then, when multiple spreadsheets and datasets are being used with many different calculations, it's easy to lose your place, which makes it difficult to find and correct a mistake you may have made several changes earlier. And when using Excel, people rarely document their various dataset versions, and version control becomes a problem. All these issues could lead to a lot of wasted time spent troubleshooting and doing additional data cleaning, making collaboration extremely difficult.

Explaining SQL/In-Database Coding

SQL is the main programming language that allows your database servers to store and edit the data on it. In database systems, SQL statements are used to generate queries from a client program to the database. This allows the users to execute a wide range of fast data manipulation in the database.

The range of functions offered within most implementations of SQL has tended, however, to fall short of the needs of someone doing data preparation beyond the need to join tables together and apply filters to slim down the amount of data to be transferred to the environment where the real analysis will be performed, usually in R or Python.

Yet many of us use SQL regularly because the data we use lives in a SQL compliant database, and if we want to do something with it, we have to write a query.

Although SQL is commonly used by engineers in software development, it's also popular with data analysts for a few reasons:

- It's semantically easy to understand and learn.
- Because it can be used to access large amounts of data directly where it's stored, analysts don't have to copy data into other applications.
- Compared to spreadsheet tools, data analysis done in SQL is easy to audit and replicate. For analysts, this means no more looking for the cell with the typo in the formula.

- SQL is great for performing the types of aggregations that you might normally do in an Excel pivot table — sums, counts, minimums and maximums, etc. — but over much larger datasets and on multiple tables at the same time.

Have a look at the disadvantages of SQL:

- **Complex interface:** Because SQL has a complex structure, it becomes difficult for certain users to access it.
- **Implementation:** Collaboration support is weak as is support for data lineage. Certain databases also implement proprietary extensions to standard SQL, which causes vendor lock-in.
- **Partial control:** Because there are certain hidden rules and conditions, the programmers who use SQL don't have power over the database.
- **Expensive:** The time and cost involved in running SQL operations daily are too high.

Using a Desktop-Only Tool

A typical desktop data preparation tool, for example, Alteryx, often takes a traditional client-server approach, with the desktop client deployed outside of the cloud, usually on-premises. Excel and certain ETL vendors fall into this category. A desktop-only tool is normally used for handling departmental-level data preparation jobs for a small number of users who require little collaboration with each other.

WARNING

The desktop-only data prep tool does come with its drawbacks:

- It isn't integrated with cloud services and can't scale with increasing data volumes as a result of that.
- When a desktop-only tool needs to deal with enterprise-scale data preparation projects in the cloud, it can't leverage the native cloud services to deliver elastic scalability and cost efficiency.

  Instead, the desktop-only solution requires a number of proprietary, separate component systems to be deployed in

the cloud to manage governance, job orchestration, execution, and sharing.

- To overcome the scale limitation, users have to either over-provision every worker node to accommodate the largest possible workload or add more infrastructure to meet the growing demand and performance requirements. Both these approaches drive up management complexity and cost.
- Desktop-only data preparation solutions don't provide agility due to its legacy waterfall design approach.
- Whenever an error occurs during the data preparation process, a user can't easily identify the root cause of the issue when there's no real-time visibility into the process.

  Instead, the user must restart the entire process in order to scrutinize all the transformation steps. Such rigid design leads to longer analytics development cycles and delayed time to results and basically means that your downstream use of the data is limited to that tool, whereas organizations have many analytics tools. A single department could have more than ten tools.

Other limitations with a desktop-only tool also include the lack of team collaboration support since this approach is focused on optimizing for one user. Another consequence of this approach is also poor data lineage control over the data life cycle.

Describing Embedded Data Preparation in an Analytics Tool

Analytics tools exist both as desktop-only solutions and cloud native solutions and many of these tools comes with a built-in data preparation capability. So, why not just use that one? Well, like any other application on the market, additional capabilities added to the main capability of the application have a tendency to offer basic functions but insufficient support. To put it simply, an analytics tool is built to first and foremost enable great analytics, not to prepare data. Of course it can offer you basic functionality, but be prepared for it to use pretty rudimentary data preparation functions.

If you're looking for some powerful, dynamic, flexible, and scalable data preparation support with the ability to scale, automate, and utilize artificial intelligence (AI) support, you won't find that embedded in an analytics solution, no matter how good the analytics solution is. Usually the analytics vendor sells the data preparation application separately with predefined and easy-to-use application programming interfaces (APIs) to the analytics solution.

Diving into a Cloud-Based Data Preparation Solution

A data preparation solution designed for the cloud is a critical component of your modern analytics and machine learning stack. With tight integration with the native cloud services, an interactive web-based user experience, as well as enterprise-class, centralized governance, and execution, a data preparation solution architected for the cloud allows organizations to explore and run a wide range of use cases at scale.

However, data in the cloud can be extremely messy, and messy data provides no value until it's cleaned up. To get data ready for analytics on cloud, companies need to take into consideration both the characteristics of the data and the use cases they want to explore in a cloud environment and select a data preparation solution designed for the cloud as part of their modern analytics stack.

REMEMBER

To address the demanding requirements for scaling, performance, and management associated with the analytic projects on cloud, the architecture of a data prep solution is crucial. The modern solution, when compared with legacy desktop-only data prep tools, follows a fundamentally different design principle.

Another aspect to consider is user experience. When expanding your company's analytics adoption in the cloud, this becomes very important. With most data now stored in cloud data lakes and data warehouses, users with various skill sets have easier access to the data without relying on IT to provision the data for them. A modern, cloud-native data preparation solution can empower all types of users, from technical to business users to easily wrangle the data in the cloud with an intuitive, modern data preparation interface.

IN THIS CHAPTER

- Introducing the fundamental parts of data preparation
- Describing the key cornerstones of data profiling
- Explaining how machine learning enhances data transformation
- Learning key principles in cleaning data to improve data quality
- Industrializing data preparation

Chapter 2
Explaining Modern Data Preparation

Data preparation, also called *data wrangling,* is the process of cleaning, structuring, and enriching raw data into a desired format for better decision-making in less time. This modern, self-service approach has three fundamental segments, which include data quality, data transformation, and data pipelines.

WARNING

Data preparation is a necessity in any company, but at the same time, the way it's approached in many companies is still not especially efficient. Despite the best efforts and intentions in most companies and organizations, it's widely acknowledged that data preparation still accounts for up to 80 percent of the effort in any data science initiative. On top of that, data has become more diverse and unstructured, which means that more time needs to be spent on removing, cleaning, and organizing data to enable any type of analysis to be made. At the same time, with an increased focus on data-driven businesses, the dependency on quality data is stressing the importance of a self-service enabled, reliable, and efficient data preparation capability. As data starts to influence just about every business decision, business users have

less time to wait for the data and require self-service capabilities in data preparation.


TIP

One way to speed up the data preparation flow is through a self-service model, which lessens the dependency on an IT-led data preparation, to a more democratized model of self-service data preparation/wrangling.

In this chapter, I introduce you to the fundamentals of the data preparation workflow and explain some of the key concepts for you to grasp in order to get your data management efforts working more efficiently.

Walking through the Data Preparation Workflow

In its simplest form, data preparation is the method of collecting, cleaning, processing, and consolidating the data for use in analysis. Simply put, it enriches the data, transforms it, and improves the accuracy of the analytical outcome. It's a step in the analytical process that consumes a significant amount of time and effort. However, too many people regard data preparation as janitorial work — as an unglamorous rite of passage before sitting down to do "real" work, meaning, for example, the task of data analytics or training a machine learning (ML) model.

The fact is, data preparation is as much a part of the data analysis process as the final results are. Data preparation, when it's properly conducted, gives you insights into the nature of your data that then allows you to ask better business questions. It may even help you ask the right questions, rather than the ones you have assumed are correct to ask.


REMEMBER

Data preparation isn't something that's done as one big step; instead, it's done iteratively. Each step in the data preparation process exposes new potential ways that the data preparation may have to be reiterated, with the main objective of generating the most robust final analysis. Figure 2-1 shows the different steps in the data preparation/wrangling process.

So, what does data preparation mean in practice? What is it that you can really expect from this activity?

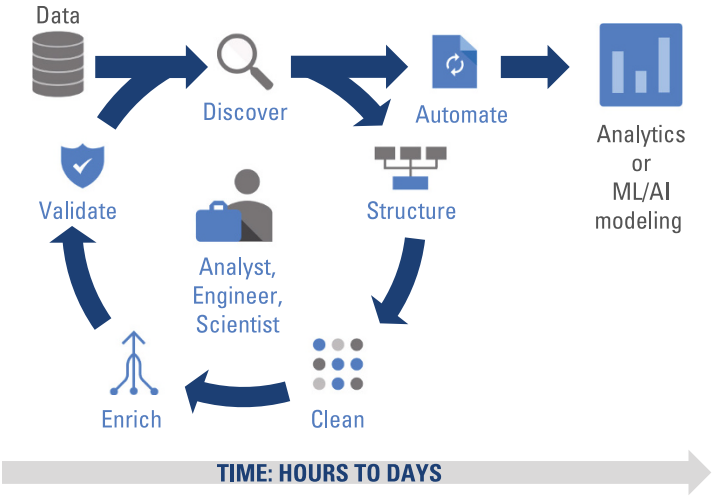

FIGURE 2-1: A typical data preparation process.

The steps in data preparation can be divided into three different areas. Each of these areas includes a set of different tasks, which I describe in the following sections.

Data quality

Data quality refers to the accuracy and cleanliness of data. It includes examining data consistency, completeness, and relevance. Reliable data quality is required for strategic decision-making when working with organizational data in an enterprise.

REMEMBER

In general, companies know that they have good quality data when they're able to use it to communicate effectively, to understand customer needs, and find effective ways to serve their customer bases. Data quality can be divided into three different parts: discover, validate, and operationalize:

- **Discover:** Before you can dive deeply into the data, make sure you understand what's in your data because it also guides you to how you want to analyze it. This step is also referred to as *data profiling*. How you wrangle customer data, for example, may be informed by where your customers are located, what they bought, or what promotions they received.

- **Validate:** Data validation rules are repetitive programming sequences that verify data consistency, quality, and security. Examples of validation include ensuring uniform distribution of attributes that should be distributed normally (for example, birth dates) or confirming accuracy of data fields through a check across the data.
- **Orchestration:** To operationalize your data quality efforts, you need to deploy your data preparation recipes in a production setting. That means automating data quality tasks as part of the data pipelines that feed analytics processes, AI workloads, and more. This is done through orchestrating a central command where you can sequence when flows run, set flow outputs, determine how alerts are sent, and more.

Data transformation

Data transformation is the process of converting data from one format to another. The most common data transformations are converting raw data into a clean and usable form, converting data types, removing duplicate data, and enriching the data to benefit an organization. Organizations may transform data to make it compatible with other types of data, move it into the appropriate database, or combine it with other crucial information. The steps of data transformation typically include

- **Structure:** Structuring data means organizing it, which is necessary because raw data comes in many different shapes and sizes. A single column may turn into several rows for easier analysis. One column may become two. Movement of data is made for easier computation and analysis.
- **Clean:** What happens when errors and outliers skew your data? You clean the data. What happens when United States individual state data is entered as CA or California or Calif.? You clean the data. Null values are changed, and standard formatting is implemented, which ultimately increases data quality. See the later section "Cleaning Data to Improve Data Quality" for more on this step.
- **Enrich:** Evaluate your data and strategize about how other additional data may augment it. You may ask yourself what new types of data can be derived from what you already have or what other information would better inform your decision-making about this current data.

Data pipelining

On top of keeping your data quality under control and managing data transformation, you also need to address the efficiency in your data pipelines. One important aspect to consider is your automation capabilities when deploying data preparation recipes directly into data pipelines that feed analytics processes and AI/ML workloads. Data pipelining involves

- **Connection:** This step refers to the connectivity framework needed to actually access the data, secure data integrations, and collect the data. A robust connectivity and API framework enables users to access live data without requiring them to pre-load or create a copy of the data separate from the source data system. This framework includes connecting to various Hadoop sources, cloud services, files (CSV, TXT, JSON, XML, and so on), and relational databases. All the connectors should support governance and security features such as roles and permissions.
- **Publication:** Data publishing is the part of the data preparation process when data is released and made available for use by the analysts or similar. It could seem like a simple enough step to take, but it's well known that the quality and speed of insight in a company relies to a large extent on the ease with which data can be accessed and utilized across an organization.
- **Operationalization:** Organizations are best served when the components of data preparation can be automated, and more readily reused and deployed into operational systems and production areas as part of the company's data pipeline design. A data preparation approach, which accounts for how the end-to-end data flow needs to work, including in a production setting, empowers the entire enterprise to make the most of all its valuable data assets in any given step of the process.

Identifying the Principles of Data Discovery and Profiling

The bulk of your data preparation work involves transforming the data itself, including activities such as manipulating the structure, granularity, accuracy, and scope of your data to better align

with your analysis goals. When you're working on a data project, you often don't have time to look at every field of every record. Discovering what's in your data, or *profiling* and exploring your data, is the activity that helps you know what's in your dataset and allows you to validate that your data transformation efforts functions as intended. Profiling your data is especially important with data that's unfamiliar to you.

REMEMBER

Data profiling is about the process of examining the data collecting statistics or informative summaries about that data. The purpose of these statistics may be to find out whether existing data can be easily used for other purposes or to improve the ability to search data by tagging it with keywords, descriptions, or assigning it to a category. But it can also be to assess the risk involved in integrating data in new applications or whether known metadata accurately describes the actual values in the source database. Data profiling also helps you identify data challenges early in the data preparation process, so late surprises causing delays and increased cost are avoided.

Often, however, data profiling is used to evaluate the quality of your data. As a frequent part of data preparation, you need to be able to quickly determine if any records contain data that may cause problems during the data transformation process.

Profiling can be done from two slightly different views:

- Examining individual values in your dataset
- Examining a summary view across multiple values in your dataset

Regardless of which of the views you use, the profiling information can be captured in text format — for example, in a list of data values, a table of summary statistics, and so on. It's also possible to build visualizations to capture profiling information about your data, or you can use data profiling tools with this capability built in.

TECHNICAL STUFF

Ultimately, individual values profiling boils down to determining the validity of individual record field values. This type of profiling comes in two forms: syntactic checks and semantic checks. Syntactic constraints focus on formatting, and semantic constraints are rooted in context. Set-based profiling attempts to determine the validity of groups or distributions of values in a particular record field.

TIP

All of these data transformations are best performed with tools that provide meaningful feedback so that the person performing the data preparation is assured that the effort was successful. In many cases, a predefined (and, hence, somewhat generic) set of profiling feedback is sufficient to determine whether an applied data transformation was successful or not. In other cases, customized profiling is required to make this determination. In either event, the bulk of data preparation involves frequent iterations between profiling, validation, transforming, and operationalizing your data.

Enhancing Data Transformation with Machine Learning

Manual data preparation, often by spreadsheet, is not only time-consuming but also often redundant. That's because different users (or even the same user) may perform the same work without necessarily generating the same results each time.

REMEMBER

Now more than ever, organizations are setting serious goals around implementing machine learning models across all areas of their business. *Machine learning* is a technique that allows computers to identify and "learn" patterns in the data, as well as identify deviations or anomalies in the data. Based on its learning, it can then perform tasks without explicit instructions. Because machine learning models replace manual programming, data scientists are able to arrive at conclusions in a fraction of the time it would've taken them. That assumes, however, that it would be possible for the model to be manually re-created, which may not be the case depending on its level of complexity.

In the early days of machine learning, it proved to be particularly useful in relation to customer behavior and fraud detection initiatives, but lately the applications for machine learning are used for all sorts of solutions, using all sorts of data and the understanding of its usefulness is virtually exploding in society.

Data preparation, whether it's for enabling machine learning or other data analysis, is essential. In the case of machine learning, where the amount of required data preparation doubles or triples in order to supply significant training data, data preparation is especially tedious.

REMEMBER

So, whether it is to reduce the time spent on data preparation for data analysis, or to accelerate the process of preparing data for machine learning, you need to have machine learning powered guidance in the actual data preparation. By using tools with ML guidance in the data preparation flow itself, the ML models will learn from every user interaction and automatically suggest the most intelligent transformation at every instance. Machine learning powered tools can also enable identification of errors, outliers, and missing data as it's detected in the data preparation.

Experience clearly shows that data analysts and data scientists are most efficient and effective when they receive immediate feedback on interactions with their data. Disruptions in the data preparation process not only slow the end-to-end preparation work, they can also be limited due to constrained end-user tools for data preparation built on frustratingly slow and inefficient workflows. Using sampling and machine learning techniques is a way to minimize or even eliminate these disruptions and deliver a fluid data wrangling experience for data at any scale.

So, how does this actually work? Take a look at a couple of examples:

- **Immediate feedback working with data at-scale increases productivity.** Users receive immediate feedback when interacting with the content of their data and are never removed from their workflows or forced to wait for processing to complete become more productive. Sampling provides a representative dataset that allows for this feedback to happen in an efficient way.
- **Enhanced performance drives better intelligence.** Like intelligent programs for Chess and Go, enhancing performance is about constantly anticipating the next moves that a data analyst might want to make when wrangling data. Machine learning allows users to explore this space instantly with higher volumes of data and faster computation than previously possible, ranking suggestions and presenting them to users with rich visualizations of potential outcomes.

Cleaning Data to Improve Data Quality

Data cleansing or data scrubbing is an important step to improve data quality. It's the process of analyzing, identifying, and correcting messy, raw data. When analyzing organizational data to make strategic decisions, you must ensure a thorough data cleansing process has been conducted. Scrubbing data is crucial to enable quality data analysis. Good analysis rests on clean data: It's as simple as that.

All too often organizations lack the attention and resources needed to perform data cleaning to influence the end result of the analysis. Inadequate data cleansing and data preparation frequently allow inaccuracies to slip through the cracks. The lack of data scrubbing leading to inaccuracies isn't the fault of the data analyst, but a symptom of a much larger problem of manual and siloed data cleansing and data preparation, or a symptom of not having the right solutions in place to secure proper data cleaning.

WARNING

Beyond the lackluster and faulty analysis, the larger issue with traditional data cleansing and preparation is the amount of time it takes. With the vast majority of the time spent scrubbing data, it's understandable why this step is sometimes skipped. However, the last thing you want to do is automate bad decisions faster based on bad data. That could have disastrous consequences for your business.

Data cleansing can be difficult, but the solution doesn't need to be. There are new approaches to data preparation in the industry that help organizations get the most value out of their data with proper data scrubbing. With visual, user-friendly interfaces, it allows non-technical users to wrangle data and scrub data of all shapes and sizes for sophisticated analysis.

The idea with this new approach is to empower non-technical or business users to do more with their data by guiding them through the process using intelligent suggestions powered by machine learning. Easy-to-use interfaces for cleaning data include using interactive interfaces that detect and remediate data quality problems like anomalies, null values, and outliers, or replace unwanted values or patterns in columns. At the end of the day, data cleansing is about finding and standardizing or removing bad data that may distort your analysis.

REMEMBER

Many companies struggle to deal with messy data that can be hard to reconcile. Sometimes the mess is due to data that has been manually entered into systems and is therefore inconsistent or incomplete. Other times, it's messy because data is coming from multiple data sources. In these situations, it's clear that traditional methods of clustering and standardizing similar values are too slow and inflexible.

Running Data Preparation in Production

After you've refined your data and begun generating valuable insights from that data, you start executing the resulting data pipelines that need to be run regularly from the ones that were sufficient as one-off analyses. It's one thing to explore data and prototype data models, but wrapping those initial outputs in a robust, maintainable framework that can automatically provide people and resources ready-to-go qualitative data is a whole other ballgame.

A solid set of initial insights often leads to statements like "We should track that measure all the time," or "We can use those predictions to expedite shipping of certain orders." The solutions to each of these statements involve *production systems* — systems that operate in a largely automated way and with a well-defined level of robustness. At a minimum, creating production data requires further optimizations to your refined data.

The task of engineering, scheduling, and monitoring that data flow, or these *data pipelines*, is to ensure optimized data is constantly ingested into regular reports and data-driven products and services is the final, but vital step of data preparation.

Optimized data is the ideal form of your data; it's designed to simplify any additional work to use the data. Specifications related to the processing and storage resources need to be applied to work with the data. These constraints often decide the structure of the data, as well as the ways in which that data is made available to the production system.

REMEMBER

Although the goal of refining data is to support the widest set of analyses as efficiently as possible, the goal of optimizing data is to robustly and efficiently support a narrow set of analyses for a certain purpose.

Building regular reports or feeding data-driven products and services requires more than just wiring the data into the report generation logic or the service providing logic. One major source of additional work comes from monitoring the flow of data and ensuring that requisite structural, temporal, scoping, and accuracy constraints remain satisfied over time.

The fact that data is flowing in these systems implies that new (or updated) data will be processed in a constantly ongoing manner. New data will eventually vary from its historical equivalents (maybe you have updated customer interaction events, or the latest week's sales data).

Within the constraints, the reporting and product/service logic must handle this variation. This deviates from exploratory analytics that can, for speed or simplicity, use logic specific to the dataset being analyzed. For production reporting and products/services, the logic must be generalized and adhered to every time.

WARNING

Common dataset variations that drive changes to the data preparation logic include

- Extensions to value ranges, such as current dates or redefining regions or customer segments
- New accuracy issues, such as previously unseen misspellings
- Record fields that have been removed or emptied for legal compliance purposes (certain information about, such as age or gender, may be redacted)
- Appearance of duplicate records or disappearance of a subset of records due to a change in customer segment names (one or more groups might be dropped)
- Additional features or columns in the dataset

TIP

You can tighten the boundary of permitted variations to exclude things like duplicate records or missing subsets of records. If so, the logic to catch and remedy these variations will likely happen in the data optimization action instead.

IN THIS CHAPTER

- Understanding why data preparation is everyone's concern
- Identifying roles involved in preparing data
- Describing the team collaboration needed in data preparation
- Applying data preparation roles using a customer example

Chapter 3
Describing Team Roles in Data Preparation

Data preparation is a key component of modern Data Operations (DataOps) that greatly benefits organizations across the world by spreading the work across teams. Each person works on components of the overall process collaboratively, and this process is referred to as *democratization*.

REMEMBER

Democratizing data preparation increases throughput and allows you to leverage the collective wisdom of the broader organization to achieve better outcomes faster. When these processes no longer are limited to IT, it can have massive impact on the business.

If return on investment (ROI) on your data is directly proportional to the number of people using it, self-service data preparation allows IT to become the data hero. IT can put its effort on streamlining the data supply chain and unleashing more data on the organization than ever before. In turn, with self-service data preparation shifting the work to the information consumers, IT organizations can focus increasingly scarce resources on data acquisition as well as broader governance issues like reuse, standardization, security, and compliance.

TIP

Shifting to self-service data preparation in your company results in faster cycle time and better insights. The people preparing the data are the ones who know how the data is being used to drive decisions.

Is Data Preparation for Anyone?

The reality for most people working with data preparation activities is that data preparation is typically work required for someone who's overall role is more focused on analysis or machine learning (ML). To address this challenge, organizations are trying to figure out how to enable different data roles with the right supporting tools and infrastructure necessary for success.

This is similar to being asked to redesign, build, and implement a new fuel system on a passenger jet while it is in the air flying. (For all data professionals out there, kudos to you for taking this on, as no pilot would ever agree to a fuel system rebuild while in-flight.)

The volume and variety of data collected by enterprises across the private and public sector are rapidly growing. This growth is outpacing the ability to staff key projects with data professionals who have the technical skills that can work effectively to help leverage data as a strategic asset.

TECHNICAL STUFF

It has been widely documented that anyone who works with data spends 80 percent of their time cleaning and transforming data. This is often accomplished by manually writing code (R, Python, and so on) to cleanse, structure, and integrate data from various source systems for use in downstream consumption for advanced analytics or business intelligence. This approach is both error prone and is the critical bottleneck in an efficient DataOps workflow.

Empowering self-service data preparation for analysts and less technical personnel means enabling the exploration, profiling, transformation, and cleansing of data in a model that allows the exploitation of project-specific data without the need for IT involvement — although it must still ensure that critical IT governance and security policies and practices are followed. This can only be accomplished by leveraging the right tool sets that enable

self-service flexibility for business users, while enforcing corporate security and governance standards. This is exactly how to succeed (with speed) through the democratization of data preparation via self-service.

Describing Roles in Data Preparation

All data initiatives, whether for ML, data visualization, or reporting, rely on clean data. That means that data preparation is essential to any data-driven organization. Increasingly, organizations are adopting new solutions to increase the accessibility of data preparation and reduce the time involved in a governed, secure manner. The role of IT or highly skilled technical teams for data is changing and now spans a variety of different users — in particular, the data analysts, who know the data best.

REMEMBER

Given that data preparation is not only a relatively new technology, but also a new process for many organizations, successful adoption of data preparation strategies requires adjusting the roles and responsibilities of team members to reap the benefits. A sound data preparation strategy requires organizations to consider how to appropriately leverage the different skills of their team. In order to increase efficiency, each role should be clearly defined and employed at the right time.

Figure 3-1 shows an example of how different roles contribute to the success of data preparation. There are five personas typically involved in data preparation. Two of these are primary end-user personas: the data analyst and the data engineer. They're a necessary part of any data preparation job. The other three are secondary personas, however, and should be seen as more important in larger data preparation implementations. They are the data scientist, the data architect, and the data executive. These roles are important players in the data prep ecosystem but wouldn't be considered necessary or irreplaceable in data preparation.

In this section, you go through in more detail a typical set of roles involved in data preparation and what their responsibilities and activities include.

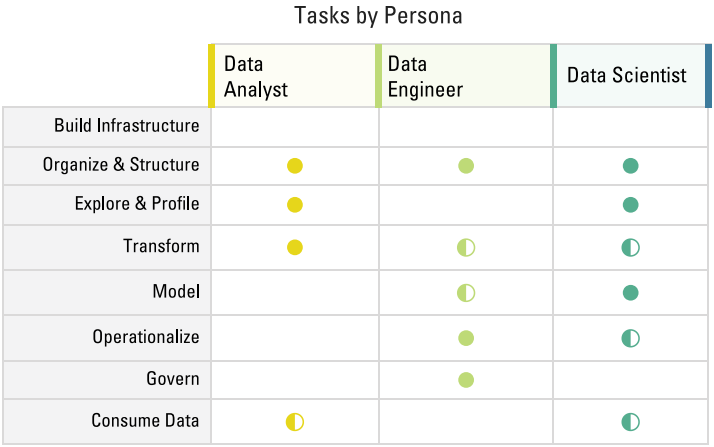

FIGURE 3-1: The roles and tasks in data preparation.

Data analysts

Data analysts deliver value to the businesses by having a deep relationship with their data. They're focused on efficiently and regularly delivering results based on knowing their data and knowing it well. Data analysts also understand the business context for the data extremely well. Perhaps better than anyone else, they know that understanding the context of your data gives you the power to answer crucial questions about your organization.

TIP

Traditionally, data analysts used to only be accountable for data reporting but are starting to also be expected to undertake data preparation and data cleansing tasks as well. With the availability of new data preparation solutions, data scientists and IT organizations are no longer completing data preparation on behalf of analysts. Instead, these self-service solutions have empowered data analysts to own the entire process end-to-end.

Data engineers

Data engineers play a growing and increasingly critical role in tying business and data preparation processes together. They are not only devoted to architecting databases and developing data pipelines (also known as *ETL processes*) but also with the somewhat unique combination of technical skills and data know-how,

data engineers can empower their more business-focused colleagues by helping them streamline and automate data-related processes.

REMEMBER

Data engineers see the bigger picture of data preparation, including scale of operations and how it fits into the business perspective. This vision makes them invaluable resources for the success of an organization's overall DataOps practices. In addition to operationalizing and building repeatable data workflows typically built by their analyst colleagues, data engineers also often provide training, scripts, and queries to help others with data preparation and analysis.

Data scientists

Data scientists combine a background in mathematics, computer science, statistical analysis, and domain knowledge to generate business value out of complex, diverse data. Data scientists typically use coding frameworks such as Python, SAS, and R to manipulate data and perform analysis.

This role is also part of the secondary personas in data preparation. While they do see value in data exploration, operationalization, and some of the data transformations that are more difficult to program, data scientists typically view data preparation tools as supplementary instead of crucial to their workflow.

REMEMBER

The goal of a data scientist is to be able to use data science techniques to engage with particularly gnarly data (for example, large volumes, complexity, a priori exploration) as well as serve as subject matter expert to other data stakeholders. Data scientists also maintain high standards for data quality and validity and push others to think critically about statistical and mathematical aspects in data preparation work.

Data architects

Data architects decide how data and the tools that access it will be configured, integrated, scaled, and governed across different organizations. The broad interest and competence of data architects mean they have a direct and important stake in any business project that uses data owned or touched by IT. Analytics initiatives need the buy-in of data architects to succeed because

they typically both govern and control the data that analysts and other stakeholders will use in these projects.

REMEMBER

Because data architects typically deal with many disparate systems and datasets, they need to understand who will use the data, how it will be used, and what the dataflow is through every system. A data architect manages the security and access controls to data sources that flow into any data preparation system. As a result, the data architect and data engineer will work closely to ensure the success of business users who are performing this data preparation.

Analytics leaders/executives

Analytics leaders understand the importance of data in delivering business value. And while they may not directly use data preparation tools themselves, they recognize how having data preparation tools deployed across their organizations leads to more efficient data pipelines, improved key performance indicators (KPIs), and potentially new insights from data. Analytics leaders can own the organizations business analytics strategy on the overall level or lead a specific department where analytics is critical, for example a marketing or finance department.

Analytics leaders appreciate tools that will make their organization smarter, faster, and more efficient, so automation and repeatable processes are crucial features that they want to spread in the organization. Analytics leaders must quickly and regularly demonstrate quantifiable value and frankly any data preparation platform that empowers their organization to own and control more of the end-to-end process is seen as a huge win.

Applying Team Collaboration in Data Preparation

The industry is starting to understand how new data platforms and applications have fundamentally changed the traditional makeup of data/analytics organizations. Furthermore, companies are also recognizing how they need to update the structure of their teams to keep up with the accelerated pace of modern business, which relies on data more and more.

If you've been reading up to this point, you know more about the most foundational roles within a modern data team and how to align skill sets to these roles, but there is more to know about how these roles work as a team.

In Figure 3-2, you can see an example of how the roles in a data team could interact during different stages of preparing the data, from raw data, to refined and finally production ready data. The work is kicked off by the data engineer who first identifies and maps all data sources needed before integrating the data sources in the data pipeline and enabling data access or data capture for the data scientists and data analysts. After they gain access to the data, they start structuring it for further data exploration and profiling by either a data scientist or the data analyst.

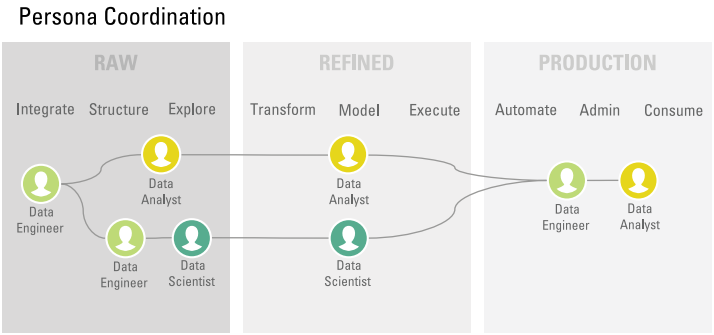

FIGURE 3-2: Team collaboration for data preparation roles.

Depending on the outcome of that step, it's decided who gets to lead the work with transforming the data for data model development. After the data is transformed, the data scientist builds and trains the data model(s) in collaboration with the data analyst with the purpose to prepare and refine the data model for production.

When the data model is ready for production, the data engineer secures that the model is operationalized, and data feeds are scheduled according to production need. After that the data architect applies security rules and allocates rights to the data as per company directives. Finally, the data is consumed by the analytics leaders for various reporting, dashboards, and business analytics or it's used further by data scientists for ML purposes.

Learning from a Customer Example

Kaiser Permanente is the nation's largest nonprofit health plan, with nearly 10 million members and 620 medical offices. Founded in 1945, Kaiser's mission is to provide high-quality, affordable healthcare services and to improve the health of its members and the communities it serves.

Kaiser wanted to radically transform the way it worked with big data by designing a data platform that would drive cutting-edge business initiatives around security, marketing, and personalized healthcare. The company mainly used claims data and provider visit data. Its specific objectives for improving data preparation were focused on

- Reducing waste — over $1 billion spent on medical supplies/equipment
- Decreasing spending — large overestimation of supply needs
- Standardizing medical procedures to ensure only a 10 percent variance in supplies required

The challenge that Kaiser had was a fragmented data architecture, which was disconnected from its data platform environment. Kaiser's business units struggled to access the data required to launch innovative initiatives, such as reducing security risks associated with HIPAA (Health Insurance Portability and Accountability Act) compliance for privacy or security of certain medical information or predicting flu outbreaks in California.

Each analyst at Kaiser used a combination of Cognos, SQL, and XLS solutions to join spend data with local procedural data in Excel. Just analyzing one procedure for one facility took eight hours, making it impossible to scale.

WARNING

Although Kaiser's IT team had invested in a powerful data platform, there was no business adoption, and the team was unable to prove return on investment. The IT organization lacked the appropriate tooling that enabled business stakeholders to access and analyze the data in the platform, while the IT team itself didn't have the resources or competence to deliver on the analytics requirements.

The solution

In the context of data prep personas, Kaiser chose Trifacta's self-service data preparation technology. With this solution, the company was able to empower non-technical business stakeholders to work directly with the data. That led Kaiser to start several key initiatives that had previously not been able to get off the ground.


TIP

Trifacta's data prep platform also helped Kaiser's IT team maintain data governance requirements for the entire organization, encouraging multiple users and use-cases while ensuring data security.

The result

Using Trifacta's data preparation platform enabled the following high-level results:

- **Faster analysis:** Analysts now deliver 97 percent faster due to the platform's ability to support self-service access, data at scale, and the automation of previously manual tasks in Excel.
- **Uncovering operational inefficiencies:** Analysts can access new combinations of financial data using Trifacta's solution. One of the first achievements was that the analysts identified a three-times disparity in spend for a common procedure across facilities, and the organization was able to improve that.

Additional to delivering analysis results faster and identifying hidden operational inefficiencies, Kaiser changed its data preparation team setup and workflow. This increased team collaboration efficiency, as depicted in Figure 3-3, and has enabled Kaiser to fully automate important parts of its data operations.

IT exercises strict control over building the architecture but collaborates with the data architect and data engineer to get it right. If the data engineer wants anything changed in the infrastructure, he works through the data architect who collaborates with IT. Data and scripts are then made available to data analysts in a "Landing Zone" where data can be picked up.

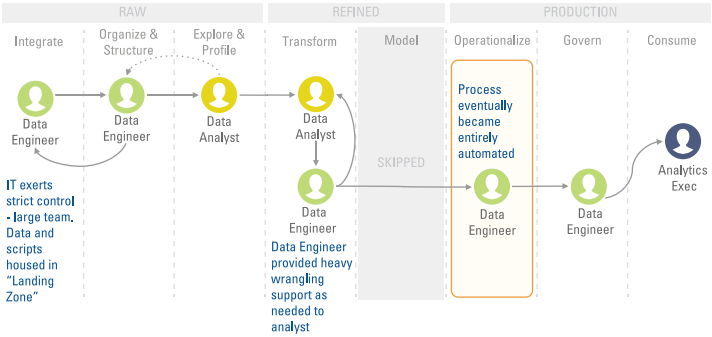

FIGURE 3-3: Kaiser Permanente's IT-governed data coordination process.

As data starts to be used by the data analysts, the data engineers provide heavy data wrangling support as needed. For Kaiser, the operationalization process eventually became entirely automated by IT and didn't require data analyst or data engineer intervention.

IN THIS CHAPTER

- Describing key cornerstones behind Trifacta's data preparation platform
- Seeing real use cases from Trifacta's clients

Chapter 4
Emphasizing the Value of Proper Data Preparation

Historically, data preparation work was hard for analysts to do themselves. Data preparation was often limited to IT, through complex coding practices that only IT could undertake. But IT didn't have the necessary business context of the data to be able to prepare the data efficiently. IT doesn't have the deep understanding needed to identify the insights and additional questions that can be explored during preparation and can help to reshape the data during the process in new and useful ways. Analysts typically defined new requirements for their IT counterparts again and again after seeing the resulting data, a cycle of unnecessary iterations between teams that can cost companies billions.

REMEMBER

User-friendly data preparation platforms are changing in a way that non-technical analysts are able to interact with the data they know best. With an intuitive interface guided by machine learning, business analysts can now prepare data themselves. The steps to prepare data for one use case may be very different from what is required for another. That's why it's so important to know the context inside and out in order to adequately prepare the data and to ultimately produce analyses based on data that is clean, suitable, and reliable. In this chapter, I show you Trifacta's data preparation platform and two use cases as examples of data preparation.

Introducing Trifacta's Data Preparation Platform

Trifacta's data preparation platform sits between data storage and processing environments and the visualization, analytics, or machine learning tools. The platform has been architected to be open and adaptable and maintains a robust connectivity and API framework, enabling users to access live data without requiring them to pre-load or create a copy of the data separate from the source data system.

The platform architecture in Figure 4-1 is divided into three parts:

- The white boxes indicate the ecosystem and extensibility parts.
- The dark grey boxes show the user experience capabilities.
- The light grey boxes show the enterprise security and governance functions.

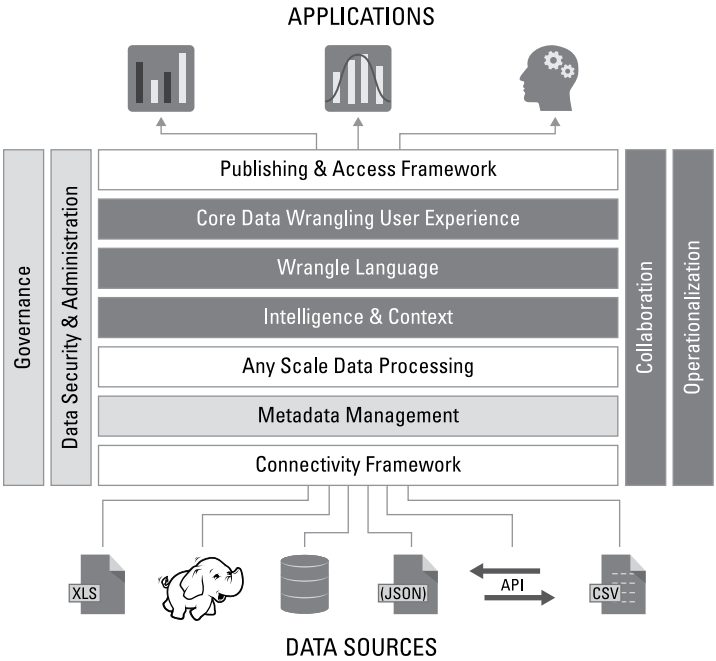

FIGURE 4-1: Trifacta data preparation platform architecture.

The architectural overview also shows how it connects to the data sources through the connectivity layer and to the application layer through the publishing and access framework.

User experience


REMEMBER

User experience is an integral part of the data preparation platform and leverages the latest techniques in data visualization, machine learning, and human-computer interaction to guide users through the process of exploring data and preparing data. Interactive exploration presents automated visualizations of data based upon its content in the most compelling profile. Predictive transformation capabilities convert every click or selection within Trifacta into a prediction and the system intelligently assesses the data at hand to recommend a ranked list of suggested transformations for users to evaluate and edit.

Trifacta also learns from data connected to the platform and how users interact with it. Common tasks are automated and users are prompted with suggestions to speed up their wrangling. The platform supports fuzzy matching, enabling end users to join datasets with non-exact matching attributes. Data registered in Trifacta are inferred to identify formats, data elements, schemas, relationships, and metadata. The platform provides visibility into the context and lineage of data — both inside and outside of Trifacta.

Core to Trifacta's differentiation is the platform's domain specific Wrangle language, which enables users to abstract the data preparation logic they're creating in the application from the underlying data processing of that logic. Advanced users can then create more complex preparation tasks including window functions and user defined functions. Every step defined in Trifacta's Wrangle language makes up a data preparation recipe or set of steps created in Trifacta that can be set into a repeatable and automated data pipeline.


TIP

In the platform, users can share reusable data preparation logic and dataset relationships, which let them leverage and build on each other's efforts. Multiple users can contribute to a single project, which parallelizes workflows, allows different degrees of participation, and speeds up time to completion. Datasets and

data preparation steps can also be integrated with third-party applications through APIs. Additionally, preparation steps can be exported and shared outside Trifacta.

The operationalization features introduce the ability for data analysts to schedule and monitor workflows that run jobs at scale in production, while still providing the traceability and access control for IT. Every data preparation recipe or set of steps created can be set into a repeatable pipeline according to hourly, daily, or weekly schedules or the time period defined by the user. Individual recipes can make up broader pipelines that make up multiple datasets and recipes.

Enterprise security and governance functions

REMEMBER

Trifacta's cloud-native data preparation solution is tightly integrated with cloud services, including storage, processing, security, and a rich set of downstream analytics services to deliver elastic scalability and security, all which are key advantages for the entire data preparation workflow.

For example, because Trifacta's data preparation platform is tightly integrated with the cloud it reads from and publishes data directly to the native storage services, such as Amazon S3, Google Cloud Storage, or Microsoft Azure Data Lake Service. For job execution, a cloud-native data prep solution uses native processing engines, such as Amazon EMR, Google Cloud Dataflow, or Azure Databricks, instead of a proprietary runtime engine to provide elastic scalability and flexibility to address the changing workload requirements.

Trifacta effectively manages data access for all users; cloud data prep uses native security policies, such as AWS IAM Role, Google Security, or Azure Active Directory, as opposed to a separate, dedicated security system. Trifacta provides end-to-end secure data access and clear auditability that comply with the stringent requirements of enterprise IT. The platform provides support for encryption, authentication, access control, and masking. Trifacta's differentiated approach to security focuses on providing

enterprise functionality (such as SSO, impersonation, roles, and permissions) while balancing extensive security framework integration with existing policies. Customers can integrate Trifacta into what's already working for them without having to support a separate security policy.


TIP

From a governance perspective there is solid enterprise data governance support integrated in the data preparation platform, which supports enriching data with geographic, demographic, census, and other common types of reference data. Common taxonomies and ontologies are automatically recognized, such as geographic and time-based content, as well as data format taxonomies for nested data structures like JSON and XML. The platform is also open/extensible through APIs, giving customers and partners the ability to seamlessly integrate additional data sources and targets.

Ecosystem and extensibility

By using Trifacta's Intelligent Execution Engine, every transformation step defined in the user interface automatically compiles into the best-fit processing framework based for the data being worked on. Trifacta transforms the data on-the-fly in the application or compiles to a variety of different at-scale processing frameworks such as Spark and Google DataFlow, or its in-memory engine: Photon. The platform offers a solid ecosystem integration with different data sources and technologies like data visualization/data science products and data catalogs. It natively supports all major cloud platforms and can handle any scale.


TIP

The Trifacta platform maintains a robust publishing and access framework. Outputs of preparation jobs are published to a variety of downstream file systems, databases, analytical tools, files, and compression formats. The system has extensible APIs and bi-directional metadata sharing with a variety of analytics, data catalog, and data governance applications. Users can share context and work between Trifacta and the external applications they're leveraging through native integration.

Learning from Data Preparation in the Real World

Even if you now know all there is to know about data preparation in theory, industrializing it and bringing it to life in your company or organization is the true challenge. Stakeholders across business and IT are always interested to learn the right way to think about applying data preparation solutions. And as with any emerging technology, the questions from organizations still learning about data preparation is often related to the actual implementation. Questions that are common and good to start include the following:

- How are other organizations preparing data, and what are the benefits they're realizing?
- Where do data preparation tools fit in the architecture?
- Who are the ideal users of data preparation technologies?
- How is security and data governance managed?

These questions and others can be answered by utilizing the use cases from two real Trifacta customers.

IQVIA

IQVIA is an American multinational company, serving the combined industries of health information technology and clinical research. The company provides biopharmaceutical development and commercial outsourcing services focused primarily on Phase I-IV clinical trials and associated laboratory and analytical services, including consulting services.

Healthcare is an industry designed to help humans and healthcare organizations that have for years attempted to improve convenience and outcomes for their patients through data. However, over the years the healthcare industry has come to realize that its analysts haven't been effective in their reporting due to the limitations that come with widespread use of spreadsheets and other limited data tools, and IQVIA hasn't been an exception from that.

IQVIA's ambition is to create a world where people in need of healthcare are empowered through their data; however, in order

for IQVIA to achieve that goal, it realized the need to entirely revamp its critical data processes.

Challenges for IQVIA

As medicine has evolved to become more precise, it has also become more difficult to make accurate assumptions. In an industry that functions on non-identifiable data, it has proven extremely difficult for IQVIA to generate a unique perspective and develop individualized solutions for its customers. With over 70 different teams and more than 2,000 people accessing data across 250 different vendor warehouses, it would take data scientists days at a time to copy data from its fragmented structure into a single system.

WARNING

Historically, data professionals in IQVIA spent 80 percent of its time preparing data and only 20 percent of the time analyzing it. This situation isn't unique in the industry. By relying on a legacy, siloed data quality process to handle the speed, scale, and diversity of its data, the IQVIA data operations disconnected from its business goals and the results the teams were able to deliver were limited.

This lack of results in data management led to IQVIA issuing a new corporate directive. The new directive was to reduce the time it took to turn around clinical trial analytics to clients by a third of the time. This objective was challenging, and as the volume of non-identified and unstructured patient data increased in IQVIA's data warehouses, it became apparent that radical changes were needed. An important first step was taken when IQVIA decided to integrate new techniques and technology to help the analysts explore and replicate datasets faster and reduce project turnaround time to meet the new corporate directive.

Solution

As part of the radical changes in data preparation that IQVIA needed to achieve, the main step was to introduce a new data management platform from Trifacta. That measure in itself resulted in no delays due to the increasing amount of data in the billions of rows of datasets that was overwhelming the company.

TIP

The increased efficiency and control of the new data management platform also helped data scientists take the time they needed to analyze data, rather than mainly prepare the data.

The new platform that enabled data visualization of each dataset through graphs, charts, and plots created a more interactive data preparation experience that allowed IQVIA's non-technical users to conduct a deeper analysis and insight on a shifting perspective. Working with non-identified patient data allowed the team to see trends more clearly and communicate those trends and opportunities to its customers. Through more straightforward reporting, its customers were able to process the data quicker to ensure the correct patients were being submitted to clinical trials sooner.

Key benefits

Prior to working with Trifacta, the manual process took about four to six weeks for IQVIA to standardize clinical study data in a format that was compliant with the Food and Drug Administration (FDA) standards. On average, IQVIA conducts 50 to 70 studies at any given point in time, and by using Trifacta's flow template, IQVIA is now able to generate a report in 15 minutes — cutting this process down to one to two days.

In total, IQVIA has reduced its overall turnaround time in data preparation by 92 percent for every new clinical study. In addition to these great improvements, IQVIA has seen an increase in prediction accuracy by four times its original rate. This has not only helped the company achieve its mission of finding the right patient for the ideal clinical trial, but also has increasingly accelerated the pace of medical discovery.

PepsiCo

PepsiCo's Collaborative Planning, Forecasting, and Replenishment (CPFR) team provides data and analyses that enable effective retail sales management. To strike the right balance between appropriate product stocking levels and razor-thin margins, PepsiCo continually aims to refine sales forecasts.

The challenge at PepsiCo

PepsiCo's customers provide them with reports that include warehouse inventory, store inventory, and point-of-sale inventory. PepsiCo combined this data with its own shipment history, production numbers, and forecast data. Each customer had its own data standards, which didn't correspond with each other or PepsiCo's system. For example, PepsiCo relied on the Universal

Product Codes (UPC) to identify each product, while customers created their own internal numbers.

WARNING

Wrangling this data involved a number of challenges:

- **Lack of standardization:** Each customer provides its data in a different file format and method. The data needs to be collected, cleaned, and standardized for analysis.
- **Reactive, not proactive:** The company was unable to deliver sales forecasts in a timely fashion for management to steer the course on sales. Not being proactive caused PepsiCo's forecast accuracy to suffer and opened the company up to lost sales and chargebacks from customers.
- **Inefficient process:** The time-consuming data preparation effort on combining retailers' data and PepsiCo supply data could take up to six months. Due to the effort to collect and prepare customers' data, analysts would only leverage this data once a month or not at all. PepsiCo Supply Chain only focused on the top ten customers.
- **Multiple platforms:** Data was spread across multiple platforms, including SAP, SQL Server, Oracle, and data received from third parties, which caused complexity, delays, and increased cost.
- **Lack of data quality control:** The company's use of Excel for analysis was error prone. PepsiCo lacked an efficient, automatic way to spot errors, which led to potentially costly outcomes.

The solution

In order to drive faster time to better forecast results, a modern cloud data preparation solution that could streamline the existing data preparation process was critical. So, to bring consistency to the data, PepsiCo turned to Trifacta.

PepsiCo selected Microsoft Azure as the cloud platform to store and process its sales data. Reports would run directly on Azure without involving multiple steps with Access and PepsiCo servers. The process would allow analysts to directly manipulate data by using Trifacta, and the adoption of Trifacta on Azure would help the team drive the business forward, increasing visibility into customer orders.

Key benefits

Gaining insight from customer data faster than ever has enhanced PepsiCo's process to offer its customers best-in-class service. It has also given PepsiCo a huge competitive advantage over other Consumer Package Goods (CPG) businesses.

TIP

The benefits from this solution included

- **Accelerated analytics:** Supply chain analysts were able to reduce the total reporting time by 70 percent and build dashboards for new customers 90 percent faster than with Excel and Access.
- **Proactive results:** Within the first few months using Trifacta, PepsiCo analysts found a $7 million order error that would have gone unnoticed.
- **Expanded reporting:** Analysts now have time to create dashboards for more customers, including online retailers.

IN THIS CHAPTER

- Moving faster when using the cloud
- Leveraging cloud-enabled ease of iteration to improve data preparation
- Enjoying the ability to scale up and down depending on need
- Collaborating efficiently
- Improving data accuracy in the cloud

Chapter 5
Ten Benefits of a Cloud Data Preparation Solution

Each *For Dummies* book ends with a Part of Tens chapter. This book is no different, so here I give you ten benefits of a data preparation solution built for the cloud:

- **Speed:** Move faster and enable more people with context for the data to get it ready for reporting and analysis.
- **Agility:** Promote agility — not a waterfall ETL process — but one where you always have eyes on the data, which in turn facilitates more exploration and iteration in your data preparation.
- **Efficiency:** Improve efficiency and TCO by leveraging a modern cloud platform, which is cheaper from a HW/SW cost perspective and doesn't require trained specialist(s) to manage the underlying hardware.
- **Scalability:** Utilize the ability to quickly and easily scale-up or scale-down resources like processing and storage capacity depending on your data preparation needs.

- **Collaboration:** Foster collaboration through a cloud-based solution that offers an intuitive user experience and facilitates the collaboration among stakeholders.
- **Quality:** Ensure quality and accuracy of the analysis and better results with an iterative process that involves multiple stakeholders.
- **Governance:** Manage data governance and data lineage in order to manage data access and have fine-grained visibility into the lineage of data transformations.
- **Innovation:** Facilitate innovation and development of new use cases and adding new data sources to your analytics on-the-fly.
- **Integration:** Enjoy seamless integration in the cloud with best-of-breed storage/processing, analytics, and data science solutions.
- **Orchestration:** Empower centralized data preparation orchestration where companies can avoid one-off data preparation and point-to-point solutions by centralizing and automating the scheduling, publishing, and operationalizing of data preparation in the cloud.

Clean data and build data pipelines faster

This book is packed with valuable information about data preparation. You discover the rise in popularity of self-service data preparation solutions and the shift from desktop-based products to cloud-native platforms. Data preparation is no longer out of reach for non-coders. With the help of this book, you find out more about Trifacta's data preparation platform and the benefits it can provide any data team.

Inside...

- Learn about the data prep bottleneck
- Explore modern data prep solutions
- See customer benefits using data prep
- Dive into team roles for data prep
- Discover cloud data prep advantages

Ulrika Jägare is Head of AI at Ericsson North America. She has a decade of experience in data, analytics, and AI/ML, as well as 20 years in telecommunications. She's the author of *Data Science Strategy For Dummies*, as well as several other *For Dummies* custom titles.

ISBN: 978-1-119-70156-9
Not for resale

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

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