LEARNING MADE EASY

Prophecy Special Edition

Low-Code Data Engineering on Databricks

for dummies
A Wiley Brand

Enable everyone to easily transform data

Boost data team productivity

Ship trusted data products fast

Brought to you by

Prophecy

databricks

Floyd Earl Smith

About Prophecy

Prophecy is on a mission to make it simpler and faster to leverage the promise of data. Through our low-code data platform, data teams of all skill-levels can visually build transformation pipelines and convert it into high-quality code, democratizing data access to meet the analytics and machine learning needs of the business. Prophecy is venture-backed and headquartered in the San Francisco Bay Area. To learn more, visit prophecy.io.

About Databricks

Databricks is a data and AI company. More than 9,000 organizations worldwide — including Comcast, Condé Nast, and over 50 percent of the Fortune 500 — rely on the Databricks Lakehouse Platform to unify their data, analytics, and AI. Databricks is headquartered in San Francisco, with offices around the globe. Founded by the original creators of Delta Lake, Apache Spark™, and MLflow, Databricks is on a mission to help data teams solve the world's toughest problems. To learn more, follow Databricks on Twitter, LinkedIn and Facebook.

Low-Code Data Engineering on Databricks

Prophecy Special Edition

by Floyd Earl Smith

for dummies®
A Wiley Brand

Low-Code Data Engineering on Databricks For Dummies®, Prophecy Special Edition

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

ISBN: 978-1-394-20592-9 (pbk); ISBN: 978-1-394-20593-6 (ebk)

Publisher's Acknowledgments

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

Project Manager: Carrie Burchfield-Leighton
Sr. Managing Editor: Rev Mengle
Acquisitions Editor: Traci Martin
Client Account Manager: Cynthia Tweed

Table of Contents

Introduction

Data analytics has undergone revolutionary change. First, easy-to-use business intelligence tools made analytics on relational data available to data users across organizations. Now, machine learning (ML) and artificial intelligence (AI) deliver value, most often by using unstructured and semi-structured data to drive industry-changing innovations from recommendation engines to automating processes.

The fuel that powers analytics and AI is data, and data is captured in various ways and formats and is delivered by data engineers for downstream use cases that enable smarter decision making and data-driven innovations. Many data sources exist, and these sources involve different types of data, such as customer data or sensor data generated by Internet of Things (IoT) devices, and more and more demands are placed on that data to create new data products and drive business results.

As demands on analytics have increased, data engineering has become a bottleneck, due to legacy technologies and resource constraints. Organizations are responding creatively by adopting the data lakehouse — a new kind of data architecture that facilitates easy access of data and fast analytics for all kinds of data: structured, semi-structured, and unstructured. Business intelligence, ML, and AI are all served from one data store.

But the complexities of data engineering still persist for many companies. Prophecy provides low-code data engineering for data lakehouses and data warehouses. Through an intuitive visual interface, all data users can perform their own data engineering to build performant and reliable data pipelines. Professional data engineers are freed from tedious and trivial work and provide real value by managing the overall flow of data throughout the organization. Business data users can self-serve with a visual pipeline builder in Prophecy to access and transform data for their needs.

About This Book

Low-Code Data Engineering on Databricks For Dummies, Prophecy Special Edition, describes the advantages of the open source lakehouse environment, pioneered by Databricks. This environment is new, introduced in 2020, but its adoption in the enterprise world is happening at blazing speed. Several sources confirm that more than half of enterprise IT shops are already using the lakehouse, with more to follow soon.

This book also introduces the Prophecy low-code platform for data engineering in the lakehouse environment. Prophecy goes beyond what past low-code software offered and serves as an open door to non-experts and a power tool for professional data engineers, all at the same time. This *For Dummies* book examines key trends in the world of data and shows how the data lake and the Prophecy platform help you surf those trends with panache. I hope you enjoy reading it as much as I enjoyed pulling it together.

Icons Used in This Book

From start to finish, this book uses icons as a visual guide to important points to remember, real-life examples, and technical considerations.

REMEMBER

The Remember icon indicates information that's worth retaining across chapters in this book and that's highly likely to be useful even when you set this book aside.

WARNING

Avoid possible missteps by paying extra attention to content within this icon.

CASE STUDY

The Case Study icon points out stories about companies using low-code data engineering on the lakehouse to save time, cut costs, improve productivity, and integrate better with existing systems.

TECHNICAL STUFF

The jargon beneath the jargon is explained.

Beyond the Book

This book can introduce you to new and improved approaches to data architecture and show you how to make data engineering a tool your entire data team — from data engineers to data analysts and data scientists — can take advantage of and contribute to. If you want resources beyond what this book offers, check out the following:

- Prophecy blog: `www.prophecy.io/blog`
- Prophecy University: `www.prophecy.io/prophecy-university`
- Request a demo: `www.prophecy.io/request-a-demo`

IN THIS CHAPTER

- Observing progress in data engineering
- Understanding how low-code approaches help
- Improving the lakehouse experience with Prophecy
- Bringing the pieces together

Chapter 1
Delivering Data the Easy Way

Data has become a hugely powerful resource for organizations, on par with human capital and financial resources. It may be a well-worn phrase, but "data is the new oil" still holds a promise that companies can put to good use.

Today, data powers everything from the most routine business operations to the most exciting breakthroughs, such as the recent emergence of ChatGPT, Dall*E, and other new platforms powered by large language models (LLMs). Capable and flexible use of data, for everything from business intelligence (BI) to machine learning (ML) and artificial intelligence (AI), is ever more critical as a business imperative to drive decision making and maintain competitive advantages. Yet it seems that the vast majority of organizations face big barriers to getting the most out of their data, whether for new or long-established uses.

This chapter describes how the world has changed and how new solutions have emerged to help organizations adapt and get the most out of these changes. The discipline that has seen the most change is data engineering; the key technology change has been the emergence of the data lakehouse, a new kind of data

architecture that combines the flexibility and cost-efficiency of data lakes with the structure and data management features of data warehouses.

While data lakehouses are powerful solutions for enabling BI, AI, and ML use cases, they can also increase the demands on the data engineering team, causing delays for downstream data users. Low-code data engineering solutions empower these data users to more easily access the data to drive analytics that help the organization accomplish its goals without over-reliance on data engineering.

You also discover in this chapter how Prophecy, a low-code data engineering solution for the data lakehouse, delivers a fully modern solution for today's data users and data engineers.

Understanding How Data Engineering Has Changed

The amount of data available to businesses is exploding, and using data effectively has become more important. Whereas structured, relational data was once the main focus of business use cases, the use of unstructured and semi-structured data is proliferating as various industries experience a rapid advancement of technology and increased digitalization. These newer data types power much of the innovation in ML and AI that holds so much promise — and in many cases, potential for competitive challenge — for today's businesses.

In order to support this level of innovation, the infrastructure that organizations use to gather, maintain, and process data — and to deliver results — is also changing rapidly. The emergence of the cloud has increased the IT capabilities available to organizations, while also increasing the complexity of managing data processing.

A rich and confusing mix of on-premises, cloud, and hybrid cloud and multi-cloud capabilities faces those who seek to make the most from data resources. The data lakehouse is a new solution that's a rapidly growing part of the data architecture landscape.

With a data lakehouse, organizations bring together data from on-premises and various cloud storage solutions into a single unifying infrastructure. Structured, semi-structured, and unstructured data coexist, as in a data lake — but data analytics and data management capabilities once only found in the data warehouse, with its carefully curated structured data, are now available in the lakehouse as well. Organizations get the best of both worlds.

Data democratization — a movement to empower business professionals with the ability to access, analyze, and interpret data without having to rely on technical experts or data engineers — began with BI. This first wave of data democratization freed analytics from being solely dependent on software engineers writing code and enabled BI users to self-serve for analytics use cases. But data users can only analyze data they can get access to, and until recently, data engineering demands have overwhelmed organizations.

Now Prophecy offers data democratization in two new dimensions. The first dimension is democratization of data access. With Prophecy, data users can now self-serve for data engineering, not just data analytics. Business users are no longer dependent on data engineers writing code and performing tasks to access and prepare data; they can use visual tools to create and orchestrate pipelines themselves.

The other new dimension is democratization of data engineering capabilities. By using Prophecy on the lakehouse, data users can access structured, semi-structured, and unstructured data in a data warehouse such as Snowflake or a lakehouse in Databricks. User can also perform the data transformations needed for BI and to power ML and AI.

Grasping the Benefits of Low-Code Approaches

Low-code approaches to data engineering have already done a great deal to empower users on pre-existing platforms. In the beginning — and continuing, in many cases, to today — compiled software code, ad hoc scripting, and great expanses of SQL code have been used to create and manage pipelines.

This unmanaged approach to data engineering hasn't scaled well to a world with more and more data types powering an ever-increasing range of use cases on an ever-expanding number of platforms. This new world includes more use cases that engage multiple platforms in a rapidly growing number of data pipelines.

So widely used data management solutions include low-code and no-code data engineering tools. These tools present a graphical user interface that allows a business user to connect data sources more easily with transformation engines and route the output to various destinations.

Low-code solutions are empowering and increase an organization's ability to make full use of their data. However, both the overall data management platforms and their low-code data engineering tools tended to share many limitations.

Data management solutions emerged in the 1990s before the emergence of the cloud. Designed and optimized for a solely on-premises world, these solutions were also optimized for structured data — this was before the emergence of NoSQL — and for BI use cases, at a time when ML and AI were very far from the mainstream.

WARNING

These solutions tend to be proprietary, expensive, and limited. They vary in their ability to handle newer data types; in their friendliness to, and scalability with, the cloud; and their ability go beyond BI to the new and more demanding world of ML and AI. Their low-code solutions inherit these same concerns. They often only work within the data management environment they're part of. A limited number output industry standard code and often don't support newer programming languages such as Python or Scala, which are widely used in data science for ML and AI.

Even newer proprietary solutions aren't well adapted to the emerging capabilities of the lakehouse model. Their limitations don't allow them to extend to this new, cloud-native, flexible, and interoperable approach.

Using Prophecy on the Lakehouse

Prophecy offers a low-code data engineering solution that enables data users to easily create and orchestrate data pipelines on the lakehouse — the modern, emerging, cloud-native infrastructure for data management. Prophecy is fully up-to-date in its capabilities for both data users and data engineers. For data users, Prophecy offers access to all the major types of data repositories. Users can mix and match from a full menu of data types and build data pipelines for use in BI, ML, and AI.

Prophecy fully supports Apache Spark, including Spark Structured Streaming, Spark Batch, and Spark SQL. Prophecy also distinguishes itself by offering three distinct advantages:

- **Full support for industry and professional coding standards in code output:** Code is delivered in Python, Scala, or SQL with dbt Core, native in each case to the underlying platform. Prophecy also uses Git for version control and allows full integration with dev, test, and prod environments and continuous integration and continuous deployment (CI/CD).
- **Full synchronization between code and the graphical interface:** Called Visual=Code, this bidirectional conversion enables data users and data engineers access to a development experience that suits their respective skill levels. Visual pipelines automatically generate open source code, allowing for troubleshooting and optimizations.
- **Full extensibility by data platform teams:** Custom visual components can be created, helping standardize against business-specific needs, and saving time and money when building and maintaining pipelines.

With Prophecy, organizations don't have to choose between ease of use and adherence to professional and industry standards. Both data users and data engineers are fully empowered to do their best work and to create assets that the organization can reuse well into the future.

Pulling It All Together

The addition of the Prophecy platform helps make the data lakehouse the gold standard for effective data management and enables the democratization of data across an organization. Companies can move faster and not break things; instead, they can increase and enhance their use of powerful and flexible open standards.

Prophecy empowers data users to a previously unattainable degree while making life easier and more productive for data engineers. As lakehouse adoption continues to grow, Prophecy is taking its place as an important platform for effective use of data by a wider and wider range of users and organizations.

IN THIS CHAPTER

- Identifying current trends
- Examining how ETL has progressed
- Looking forward to what's new

Chapter 2
Updating the Data Engineering Landscape

Data engineers build systems that bring data in from the outside world and make it usable within an organization. They have a wide range of responsibilities, including data security, data management, data governance, and making sure that data is usable for many purposes. The two main types of stakeholders for the work of data engineers are business intelligence (BI) users, who largely use data to help run and grow the business, and data scientists who develop predictive analytics, machine learning (ML) models, and artificial intelligence (AI) applications. Both groups contribute to internal and customer-facing applications that make a crucial difference in business success.

The role of data engineers is changing rapidly for reasons I describe in this chapter. The increasing importance of data to organizations and the emergence of new technologies give data engineers new tools but make the role ever more complex. This is exacerbated by the fact that organizations are largely in the middle of a transition from older, largely on-premises technologies to newer, largely cloud-based technologies.

Many data engineers deal with both the old and new technology sets in their daily work, along with the ongoing need to migrate selected workloads from one to the other. Eventually, the migration will near completion, but for most organizations, that's still years in the future.

Data engineering is so important to organizations today that it's extremely difficult to hire and retain the talented individuals who do this work. The value of a data engineer increases with experience, especially experience with emerging technologies such as those I describe in this chapter.

Catching Up on Current Trends

Almost everything about data has changed in the last few decades. Both the business imperatives around data and the technology framework in which data is managed and used have changed. You discover the most important trends in this section.

Identifying business imperatives

Business and, for some organizations, research needs in a changing world have led to a few business drivers that are critical to organizational effectiveness. Key business imperatives include

- Huge growth in available and diverse data
- Rapid increase in the importance of data
- The growing importance of analytics
- The democratization of data

There's more and more data

The amount of data available is growing rapidly. Sensors are cheap and ubiquitous. New cars, especially electronic vehicles (EVs), contain multiple computer systems on wheels. Phones are smart. Watches are smart. Doorbells, for goodness' sake, are smart.

The single change that best sums up the huge proliferation of data is the move to 5G technology for smartphones. 5G transfers data much faster than previous technologies. The move to 5G is still in the early stages, but already the average 5G user consumes

about twice as much data as users of previous technologies. As 5G becomes the norm, applications will grow more powerful, moving and returning more and more data.

REMEMBER

In total, data created, captured, copied, and consumed worldwide doubles roughly every two years. Companies have to choose what data to capture and use, but the answer is always "more." As shown in Figure 2-1, data stored is growing by nearly 20 percent a year, and the requirement for data availability after it's captured is increasingly "immediately."

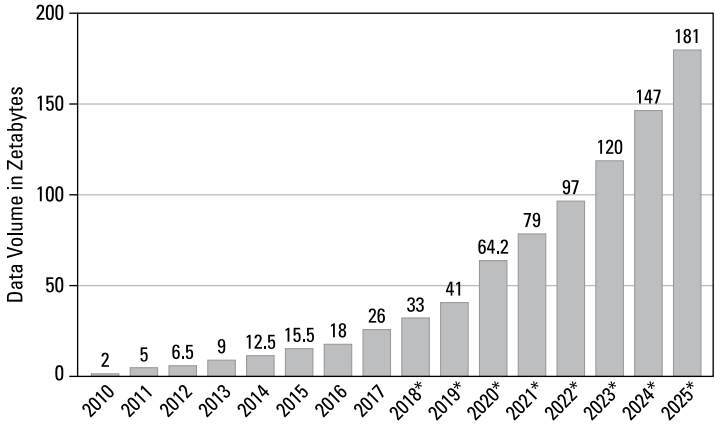


FIGURE 2-1: Growth in data creation from 2010 to 2025 (estimated after 2020).

Data is more and more important

Businesses are increasingly valued on their creative and effective use of data. Four of the top companies by market capitalization — Apple, Microsoft, Alphabet (Google), and Amazon — are largely defined by their use of data and their role in data infrastructure worldwide. Many other rising stars of the last few decades — such Netflix, Airbnb, and Tesla — are significant innovators in data use.

This trend is accelerating further today with the growing importance of large language models (LLMs) such as ChatGPT and Dall*E. These tools work by processing immense amounts of data at a cost of hundreds of millions of dollars. Companies that had not heard of LLMs a couple of years ago now have to have LLM — and, behind that, ML and AI — strategies to enable a multitude of solution from AI Chatbots to automated content generation. And

these strategies are entirely dependent on having skilled engineers to capture the right data and put it to use.

Analytics are becoming more important

As the availability and importance of data have grown, businesses have put more resources into analyzing data and making it available for internal and external use. Large companies have growing IT operations that are busily bringing in and processing data. Initially, this was mostly for transactions, such as billing, but today it's largely for analytics. (Billing data, for example, is critical for bringing in revenue, but it's also a rich resource for analytics.)

Initially, analytics were about what already happened; how much money did we make? How much do we owe in taxes for last year? But increasingly, analytics has become predictive: If we open a new retail store, where's the best place to put it? How much money will it make? Even ML and AI can be seen as a sophisticated and responsive way to do predictive analytics.

Data is being democratized

Access to data and the tools needed to use it has become increasingly democratized. Originally, business analysts learned structured query language (SQL) so they could run reports on data in transactional systems. Over time, BI tools such as Looker, Microsoft Power BI, and Tableau provided a no-code front end to this data, although savvy users could still inject SQL for improved efficiency and control. Today, these tools are increasingly used for big data (unstructured and semi-structured data) as well as transactional and other structured data for business reporting.

Finding technological drivers

Technical people create new ongoing approaches — sometimes in response to business needs, other times simply because an approach is interesting or efficient. Some of these new approaches become important trends because they meet a wide variety of needs. Key technological drivers include

- The increasing use of unstructured and semi-structured data
- The rise of open source software and data standards

» The increasing movement of data storage and data processing to the cloud

Unstructured and semi-structured data is now critical

In the previous century, most business data was structured — organized into predefined rows and columns. From the 1970s on, relational databases such as Oracle, IBM DB2, and Microsoft SQL Server grew in importance. The most important uses of data were for data records — customer information, sales data, employee data. Governments kept census data, driver's license information, and tax records.

In the last few decades, though, there's increasing need to store unstructured and semi-structured data such as images, movies, music, log files, key-value stores, and much more. To meet this need, NoSQL databases were created and have grown in importance. Google was an early leader in storing what were considered, at the time, incredible amounts of data in NoSQL databases.

Open source standards rule

The relational data business was largely proprietary, defined by large license fees paid to companies such as Oracle, IBM, and Microsoft in databases and Ab Initio and Informatica in extract, transform, and load (ETL) processing and other aspects of data management. Much of the move to NoSQL was driven by the need of companies like Google and Netflix to avoid mortgaging their futures to pay for expensive, established databases.

Infrastructure is moving to the cloud

The rise of cloud companies that deliver software as a service (SaaS), infrastructure as a service (IaaS), and other functionality over the internet was marked by the founding of Salesforce, the first highly successful SaaS company in 1999, and the launch of AWS, the market leader in IaaS, in 2006.

REMEMBER

These trends all contribute to each other. Companies started using unstructured and semi-structured data more as robust open source standards for it emerged; today's LLMs are trained on massive amounts of data stored in the cloud and could never have achieved the same power if they depended on on-premises data.

Enumerating Key Technologies

What are the key technologies that have emerged to leverage all these trends? Organizations have to ingest, process, and analyze data. These steps are organized into data pipelines, built and managed by data engineers. (Who are also responsible for cleansing, organizing, and securing data, among other important processes.)

The role of ETL in data pipelines is crucial, and it has evolved:

- In the 1980s and 1990s, ad hoc scripting was used for ETL. This get-it-done approach solved problems quickly, but the resulting scripts were error prone and hard to manage. They were also often dependent on the individual who wrote them, which caused problems as people moved on from a company.
- In the 1990s, new, proprietary approaches brought new sophistication to ETL, including visual tools for building data pipelines. ETL became central to a market in data management worth billions of dollars a year. Alteryx, Informatica, IBM, and Oracle have been among the key providers.

  These providers initially worked with relational data used by business analysts. But they've increasingly moved to support unstructured and semi-structured data used by data scientists for predictive analytics, ML, and AI, too.

In the last 20 years, the cloud has brought two new kinds of solutions, generally used together:

- The first is the data warehouse, with Teradata on-premises, Snowflake in the cloud, and cloud provider offerings (Amazon Redshift, Azure Synapse, and Google BigQuery) playing a leading role. These offerings mostly handle structured data.
- The second is the data lake, with Spark as the driving open source technology. The leading providers are Ab Initio on-premises and Databricks in the cloud, alongside managed Hadoop + Spark offerings from each cloud provider. Data lakes have mostly handled unstructured and semi-structured data, as well as structured data that hasn't yet been cleansed and otherwise made ready for use in the data warehouse.

Identifying Gaps in Previous Techniques

Initially, organizations tended to keep their data lakes and their data warehouses separate. The real-time analytics, ML, and AI teams had their NoSQL databases and Python code; the BI people had relational databases and visualization tools like Looker, Microsoft Power BI, and Tableau.

In many organizations, the separate data infrastructures have merged into a blended, two-tiered approach, as shown in Figure 2-2. There are separate pathways for less-structured data (the data lake), which are mostly used by data scientists, and cleansed, structured data (the data warehouse), which is mostly used by business analysts. Data engineers provide the infrastructure that serves both.

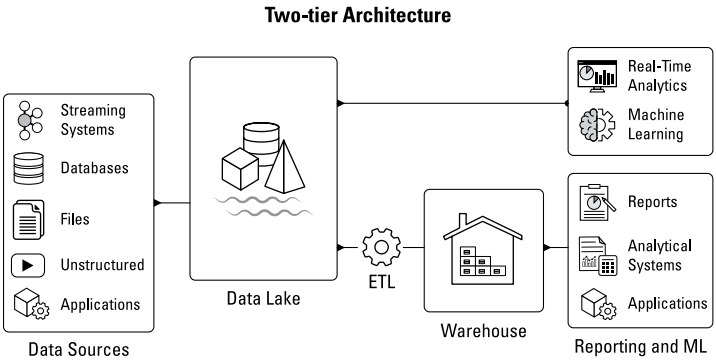


FIGURE 2-2: Many organizations use a two-tiered data processing infrastructure.

In the two-tier approach, everything goes into the data lake first. Real-time analytics, ML, and AI have access to all the organization's data, even if it's in somewhat rough form.

Then, ETL and other processes are used to prepare some data, much of it originating in transactional systems and other systems of record, for the data warehouse. This structured data is used for reporting, BI, and applications, many of them internal.

WARNING

What could possibly go wrong? Well, this two-headed beast, like most such, has some challenges. They include

- **Complexity:** Figure 2-2 doesn't show that both the data lake and the data warehouse are each often split between on-premises and cloud components, with multiple ETL and analytics processes required to try to bridge the gap.
- **Cost:** Doing the same thing two different ways often costs more. The use of multiple platforms with different tool sets adds to costs.
- **Legacy burdens:** Much of the reason for this divided approach is to allow continued use of older, expensive standards such as expensive proprietary databases and ETL tools such as Ab Initio, often running on-premises. This raises costs and traps key workloads in the less flexible on-premises world.
- **Lack of data democratization:** Complex architectures are overwhelming to those who want to participate in the democratization of data. Users end up waiting on over-worked data engineers to make data spread across multiple systems accessible to them, and the resulting solutions are often error-prone and unstable.

Examining What's New for Data Engineering

The new approach that is taking the data processing world by storm is the use of lakehouses. The term is a combination of the names of the two formerly separate repositories, the data lake (for all data, in all forms) and the data warehouse (used mostly for structured data).

The lakehouse was introduced by Databricks, a company founded by the creators of the Spark programming language in 2013. Databricks has been successful in helping companies move on-premises big data projects, many of which didn't meet their original goals, into more efficient and more successful cloud data lake projects by using Spark.

REMEMBER

The lakehouse architecture was popularized in 2020 with the launch of Delta Lake by Databricks. Originally created to make data lakes more reliable, Delta Lake has quickly added data warehouse-type capabilities to a single repository that can hold all of a company's data.

Within just a few years of the launch of Delta Lake, two-thirds of organizations surveyed by Databricks are using a data lakehouse. With a lakehouse, organizations can unify their data architecture, as shown in Figure 2-3. Data scientists still use it as a data lake, whereas BI users connect to the data warehouse-type capabilities to meet their needs. The work of data engineers is vastly simplified as more and more of the data they must prepare and deliver lives in a single repository.

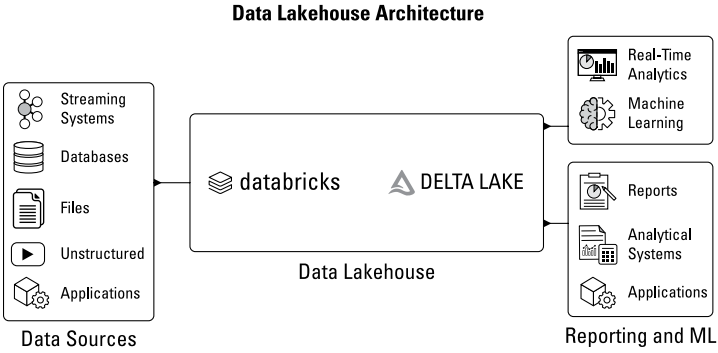


FIGURE 2-3: The data lakehouse offers a unified architecture.

The lakehouse concept has rapidly gained traction — and just in time, with 80 percent of worldwide data expected to be unstructured by 2025, according to IDC. The lakehouse handles unstructured data for ML and AI smoothly. Also, the lakehouse increasingly makes efficient processing for reporting and analytics, originally used only with structured data, available for all the information that an organization holds.

Some projects still go to a data warehouse for use with existing tools that haven't yet been brought up to date with the new approach. But the trend is to do more work directly in the lakehouse. The lakehouse is fundamentally changing the way organizations work with data and serves as a foundation for rapid progress.

IN THIS CHAPTER

- Finding challenges in coded solutions
- Experiencing benefits in low-code data engineering
- Making data transformations available to all

Chapter 3
Using Low-Code Data Engineering

Data engineers, like software engineers, do their work by writing code. Software code is powerful and runs efficiently, so it's the right tool for many jobs. However, organizations are dealing with data flows that are doubling every couple of years. They're also dealing with more complex demands, such as moving to the cloud and supporting business data users, all while battling resource constraints and an expanding ecosystem of data sources and tools.

In such an environment, manually coding for routine data engineering work can't scale fast enough to meet the challenges organizations face. The solution is low-code data engineering tools. These tools allow a much wider range of people to accomplish routine data engineering work, speeding the organization's progress and allowing core data engineering professionals to focus on the toughest problems.

THE DATABRICKS CHALLENGE

For many years, data engineers focused first on structured data and business intelligence (BI) needs. Advanced analytics, machine learning (ML), and artificial intelligence (AI), all of which required unstructured and semi-structured data to solve problems, were left to data scientists to figure out.

The success of Databricks in the cloud has opened up many doors for organizations, including the ability to meet the challenge posed by the growth of ML and AI — in particular large language models such as ChatGPT and their many potential applications.

Consequently, data engineers are increasingly including all these new types of data and new use cases as part of their core skill sets. However, it takes years of technical and programming experience to become truly skilled with new tools and new approaches. It will likely take many years for data engineering as a discipline to fully come to grips with these new demands; until then; there's an even stronger bottleneck when it comes to finding people with the skills that organizations need.

Experiencing the Challenges of Coded Solutions

It makes sense that data engineers work in code. Software code is the most powerful tool for solving all kinds of problems in data processing. When used appropriately, coded solutions enable professionals to work at a higher level and solve problems quickly.

WARNING

The problem comes when the need for coded solutions scales faster than anyone can manage. And this gap between organizational needs and available resources has grown rapidly in recent years. Check out Chapter 2 for the key drivers of these challenges. In addition to the impact of these changes on the entire organization, software engineering itself is changing in response to these and other challenges, adding even more complexity in using code as the only answer to data engineering needs.

Data engineers today use many tools to build data pipelines and perform all the supporting tasks needed to make data fully useful

to, and usable by, the organization. Chief among these are the programming languages Java, Python, Scala, and SQL.

Figure 3-1 shows how coded solutions fit into a data lakehouse architecture. While the core repository has been unified, with the lakehouse supporting all kinds of data and the full range of business needs, the coding area is still a weak point. SQL queries, which can range up to thousands of lines long and be hard to maintain, coexist with extract, transform, load (ETL) notebooks and orchestration scripts.

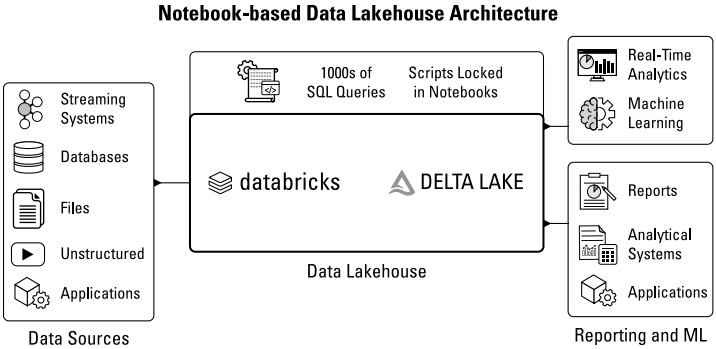


FIGURE 3-1: Scripted solutions can fall short of data engineering requirements.

There is some controversy as to whether SQL is truly a programming language. For the purposes of the book (and in my opinion), it is. It's unusually accessible, and many business intelligence (BI) tools allow users to add small sections of SQL code to extend their visual frameworks and get a lot done, quickly. But the accessibility and usefulness of SQL don't mean that it's not a "real" programming language.

WARNING

The major challenges in using coded solutions for all data engineering challenges in today's fast-changing data processing and analytics environment can be summed up as follows:

- Data engineers are overwhelmed by ever-growing demands in an environment that only gets more complex.
- Data engineers are challenged to work in multiple languages for different needs, such as Java, Scala, and SQL, requiring other data professionals to handle the different skills needed.

LAKEHOUSE AS PART OF THE SOLUTION

The move to lakehouse is part of the solution to the problems that data engineers face. Currently, data repositories are split across cloud and on-premises databases. By unifying a large and growing share of the organization's data in a single cloud repository that meets both data science and BI needs, the complexity of the data engineering challenge is reduced. Pipelines become simpler and code reuse increases.

The benefits of the lakehouse also present an opportunity to do even more with data. There will never be too many data engineers, so low-code data engineering solutions for the lakehouse are needed to take advantage of this opportunity. I describe the Prophecy offering, and some of what it can make possible for your organization, in Chapter 4.

- Existing code becomes outdated as the environment gets more complex and software standards evolve, saddling data engineers with technical debt.
- Business data users are blocked on the data engineering team for both minor changes, such as extending a table by a few columns, and for major projects.
- The organization finds it harder and harder to take advantage of new data streams and meet new business opportunities, making it harder and more time-consuming to generate business value.

Observing the Benefits of Low-Code

Over many years, existing platforms have gained low-code tools that help make the creation of data pipelines easier and more productive. BI tools such as Looker, Power BI, and Tableau offer a visual interface that puts the power of data analytics in the hands of data users, which is a big step toward the democratization of data.

REMEMBER

Low-code tools for ETL put the power of data engineering in the hands of data users across the organization. Low-code tools also make life easier for data engineers, in three important ways:

- The data engineers are spared a lot of small requests that take time to understand, implement, test, and hand over.
- The data engineers can focus their time on extending solutions created with low-code tools where needed to achieve the best results.
- The low-code tools also make many routine tasks easier for the data engineers themselves, speeding their work and reducing the potential for typing errors and other trivial, but impactful problems.

Existing low-code ETL tools, however, were created for use with previous generations of technology. They have some long-standing faults that limit their usefulness in the data lakehouse environment:

- **Structured data orientation:** These tools were initially developed when structured data was the mainstream focus. Some of them have been extended to work with unstructured and semi-structured data — but where the capability exists, it may not be fully functional.
- **BI orientation:** Data engineering teams tend to prioritize BI at the expense of ML and AI, which have only recently become mainstream business tools. So older low-code ETL tools may not fully support these newer applications.
- **On-premises focus:** Legacy tools were developed for the on-premises world, and many aren't fully cloud-capable. They aren't scalable in a cloud environment.
- **Limited orchestration support:** Creating data pipelines is one thing; orchestrating them is an additional feature set. Not all legacy tools support this well.
- **Limited lakehouse support:** The lakehouse concept is quickly moving into the mainstream, but it's only a few years old. Existing low-code ETL tools may be decades old and not fully up to speed with either data lakes or lakehouses.
- **Proprietary code generation:** Many low-code ETL tools only work within the proprietary environment in which they're offered. Code portability is limited if not completely restricted, and you can't extend the pipelines that users create outside of that environment.
- **Limited code generation:** Where a low-code ETL tool does generate non-proprietary code, that code is unlikely to fully support modern data engineering languages such as Scala, Python, and SQL.

One existing tool worth mentioning in this context is Alteryx, a flexible low-code data engineering tool that works on a laptop or desktop computer. Unfortunately, it must pull the entire dataset onto the machine it's running on to work with it. If the dataset doesn't fit on that computer, as is often the case in today's world, or if performance is unacceptably slow, the user is out of luck.

Democratizing Data Transformations

The democratization of data is further extended using low-code data engineering tools:

- BI tools make the power of data analytics available to data users, not just software engineers.
- Low-code data engineering tools make the power of data engineering available to these same data users, not just data engineers.

In both cases, the democratization of data is also good for the technical professionals. They can focus their time on the most challenging work, and they can use the same tools as everyone else to be more productive in routine tasks.

As with BI, data users are plugged into the business needs they're trying to meet. They can become skilled in the use of low-code data engineering tools. They gain the skill needed to get the most out of them and even, in some cases, to extend them with code.

With the right tools, data users can also quickly come up to speed on using ML and AI to meet business use cases, just as they've already done with BI.

Finally, the availability of low-code data engineering tools raises the quality of the dialogue between data users and data engineers. The data users handle everything that they can themselves and make informed, focused requests for data engineering help where needed. They retain a sense of ownership of the data pipelines they've created for the tasks they're seeking to accomplish.

IN THIS CHAPTER

- Getting the most from key capabilities
- Making pipelines better and more accessible
- Investigating customer applications

Chapter 4
Using Prophecy for Modern Data Engineering

Prophecy stands at the intersection of two trends that, when combined, can help solve some of the biggest problems in data management. The first is the move by organizations to the lakehouse. The second is the introduction of low-code tools to make data engineering easier across a variety of platforms.

Prophecy is a leading low-code tool for self-service data engineering for the lakehouse and data warehouses. In addition, Prophecy has added new capabilities that make it more flexible, more powerful, and more useful than low-code data engineering tools on other platforms.

REMEMBER

With Prophecy, self-service data engineering is a first-class citizen in the data engineering toolkit. With Prophecy's capabilities, there's no drop-off in code quality, manageability, flexibility, or extensibility between hand coding and code generation from the tool. The appropriate coding language is used for the platform where data resides and code can execute where data lives, reducing costs and improving performance.

Figure 4-1 shows how Prophecy fits into the data lakehouse architecture of the modern lakehouse. Data engineering gets a polished and highly usable toolkit for use with the lakehouse.

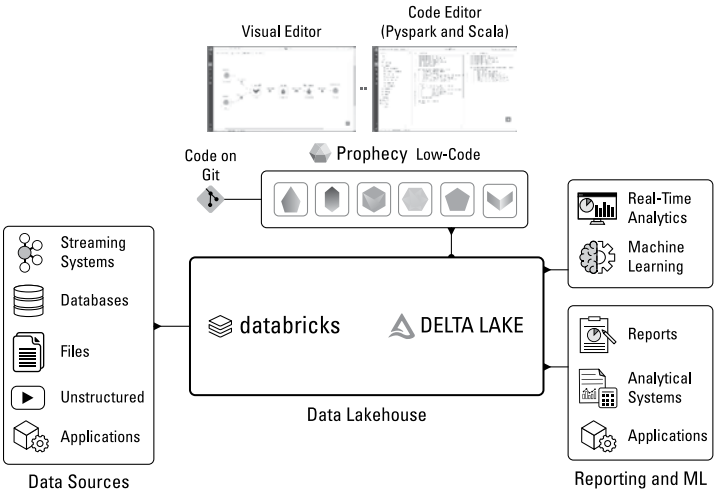


FIGURE 4-1: Prophecy offers a data engineering solution for the lakehouse.

Prophecy also offers full bi-directional synchronization between hand-coding and code created and managed in the Prophecy graphical interface. The same tool that opens up data engineering capabilities to data users becomes a major part of the data engineer's toolkit.

In this chapter, you get a brief description of how to use Prophecy's capabilities, and I share a few industry use cases.

Leveraging Key Platform Capabilities

Prophecy enables data users to visually build data pipelines for Apache Spark and SQL on the data lakehouse. The pipelines are converted by Prophecy into 100 percent open source code in Python, Scala, or SQL.

Prophecy offers the capabilities that data users need:

- Self-serve data transformation
- Sharing of data products
- Scheduling of data pipelines

The pipelines created by Prophecy are comparable to those created by data engineers, and Prophecy can be used directly by data engineers to increase the speed and reliability of parts of their work. When a pipeline is created in Prophecy, it doesn't have to be rewritten or reviewed before going to production.

Data pipelines are assembled using visual components called *Gems*, as shown in Figure 4-2. The user drags and drops Gems into place. Prophecy also offers an AI-assistant that empowers anyone to visually build data pipelines with plain English queries.

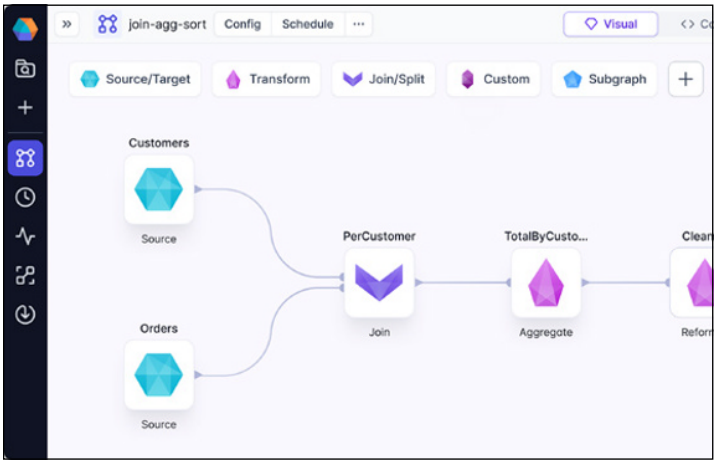


FIGURE 4-2: Prophecy users drag and drop Gems to build data pipelines.

Among other capabilities, Gems can

- Read and write common data formats such as files, databases, Kafka streams, and Delta Lake storage.
- Transform data using a wide range of capabilities including joins, filters, and aggregations.

- Create standardized templates in Prophecy's Framework Builder.
- Support Spark Batch and Spark Streaming pipelines.

After a pipeline is built, the user can run the pipeline and review the state of the data after each step. Abstractions such as subgraphs and configurations make it easy to reuse sections of the pipeline. It's also easy to develop and run tests for all components of the pipeline.

After pipelines are created, you need an orchestrator to schedule pipeline execution and alert you to failures. Two widely used orchestrators are supported:

- For Apache Airflow, Prophecy converts the pipeline into Python code. You can use it in existing Airflow jobs.
- For Databricks Workflows, Prophecy uses the Databricks API.

After an orchestration job is deployed, the user accesses the monitoring screen to check the status and see logs of job execution. These visual assets are turned into high-quality code that's fully open source with no proprietary components. For DataOps, the user connects Prophecy to a Git provider such as GitHub, Gitlab, or Bitbucket.

Managing Pipelines with Prophecy

Prophecy makes common use cases in data management much easier. For instance, within the lakehouse data is often moved through three levels of tables:

- **Bronze:** Bronze tables hold raw, unprocessed data, such as event data to be used for machine learning (ML) or retail transactions.
- **Silver:** Silver tables hold clean data that's been transformed — for instance, the features in a ML feature store or sales data for products.
- **Gold:** Gold tables hold aggregated data that downstream customers can use directly, such as the predictions of a ML model or a sales report.

In creating a reporting pipeline, the user can create three Subgraph Gems, each of which incorporates several pipelines steps. For instance, the user may create IngestBronze, MergeToSilver, and AggregateToGold Subgraph Gems to handle each step in the process.

The IngestBronze Subgraph Gem includes moving orders from their original source to the raw_orders_bronze table, including the order ID, the customer ID, and other order details. Figure 4-3 shows this step in the pipeline.

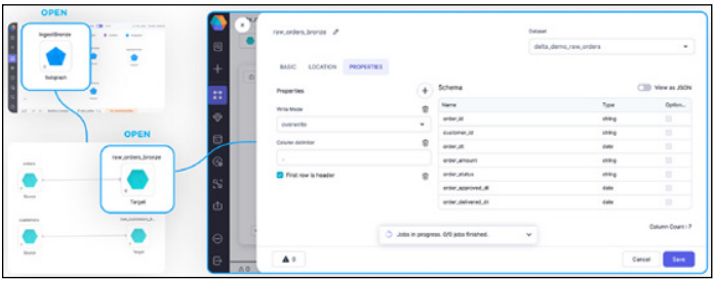


FIGURE 4-3: In Prophecy, the user can create a Subgraph Gem to handle creating a bronze-level table of raw data.

Additional Gems handle the steps needed to cleanse the source data (MergeToSilver) and create a business report (AggregateToGold).

IN THIS CHAPTER

- Making data engineering accessible in healthcare
- Benefitting from better data insights
- Increasing employee productivity in asset management
- Improving productivity
- Gaining competitive advantage

Chapter 5
Diving into Industry Use Cases for Prophecy

Prophecy customers are achieving outstanding results across a range of customer use cases by modernizing and accelerating extract, transform, load (ETL) processes on the lakehouse. This chapter gives you several examples of successful use cases.

Democratizing Data Transformations in Healthcare

CASE STUDY

HealthVerity is a United States healthcare company with more than 250 employees. It supports organizations in the pharmaceutical industry, government, and the insurance industry in connecting and exchanging real-world data (RWD), which creates an ecosystem for healthcare data.

Data and efficient processing of and access to data are at the core of what HealthVerity offers. However, the company's data engineers were tied up creating and maintaining pipelines. Non-technical

users were often blocked on projects while waiting for data engineering support.

In addition, HealthVerity needed to work with massive amounts of data in a wide range of data sources and types, including clinical data, lab data, claims data, RWD, electronic medical records, healthcare customer relationship management (CRM) data, and many others. Faced with a lack of data engineering resources, the organization was taking 8 to 12 weeks to onboard new customers, largely because of delays in getting the necessary data pipelines built.

HealthVerity adopted Prophecy primarily for use by the data engineering team and lines of business. The company had previously implemented the Databricks Lakehouse architecture. By adopting Prophecy, HealthVerity makes data engineering accessible to a much wider range of personnel. Data users are able to build and modify pipelines themselves, and data engineers have moved up the value chain, optimizing pipelines and extending the Prophecy visual user interface to include all the capabilities that data users need to be productive.

Final implementation results include the following statistics:

- Pipeline development is 12 times faster.
- Team productivity increased by 7 times.
- The company enjoyed a lower total cost of ownership of 66 percent.

As technology advances and customers come up with new requirements, HealthVerity is able to adapt and rapidly improve their offering for all customers with their use of Prophecy.

Accelerating Insights on the Data Lake

CASE STUDY

A leading Fortune 500 financial services company with more than 25,000 employees supports a payment network with global reach, supporting two billion transactions a day. Transactions must be received, checked for potential fraud, and approved within seconds. Data from the transactions is vital to the success and growth of the company and its credit card business.

The company's transaction volumes, which were already high, were growing rapidly. The organization's on-premises ETL solutions were at risk of becoming overwhelmed. In addition, the use of proprietary ETL solutions locked the company out of the rapid innovation fostered by open source. This included an inability to integrate and interoperate with other internal systems.

The company moved to a data lake architecture, using manual processes. The new architecture enabled the organization to move to infrastructure as code, providing a great deal of flexibility. However, along with all the freedom, it was now harder to implement and maintain standard approaches across the organization. Productivity also suffered.

This financial services company adopted Prophecy primarily for use by the data engineering team and lines of business. The use of Prophecy accelerated the move from Ab Initio, the ETL tool used in their previous, on-premises systems. Prophecy's transpiler generated standardized, high-quality code for use in the new environment.

With the Prophecy implementation, the company reaped the following benefits:

- Data transformations happened 5 times faster.
- Deployment times decreased by 85 percent.
- Version control and collaboration improved due to Git and CI/CD integration.

After migration, additional benefits included visual development in Prophecy helping data users be productive in the new environment. With Prophecy in the new environment, developing new model features shrunk from two weeks to a few days. Processing time has decreased, with one complicated workflow shortened from nearly three hours to less than ten minutes. ETL pipelines are standardized across the company, and the new system has become a single source of truth that the entire organization can understand and trust.

Empowering Business Users in Asset Management

CASE STUDY

Waterfall Asset Management, an investment management firm that operates globally, has more than $11 billion of assets under management. It makes critical investment decisions for its clients around the clock. Delivering outstanding portfolio performance and interacting productively with clients are vital to the continued success of the business.

The flow of available investment information continues to increase, providing both an opportunity and a challenge to firms across financial services. Waterfall was using manual processes and a legacy ETL system that couldn't handle the velocity and variety of data arriving at scale. The organization hired more engineers, but workflows slowed and data quality became a concern.

Data users with critical requests for data access and transformation were left to wait for assistance from overburdened data engineers or do without, and were obligated to spend time performing manual data work. Client service and even portfolio performance suffered.

By moving to Prophecy's low-code data engineering platform on the Databricks Lakehouse, Waterfall increased employee productivity and improved customer satisfaction. Data engineers work directly with the 100-percent open source Spark code generated by Prophecy to maintain best practices and make sure that data products are fast and reliable.

The move resulted in the following:

- DataOps productivity improved by 14 times.
- Time-to-insight for trade desk analysts increased by 4 times.
- The company saw a much faster and smoother path to better investments.

At Waterfall, data now moves at the speed of the market, improving the organization's ability to access the best investment opportunities. Anyone on the team can go from data to insights much faster, without having to speak to a team of engineers.

The repeatable frameworks provided by Prophecy make it easy for data pipelines to be standardized and shared across teams, saving time and reducing errors. Data engineering can now focus on high-value tasks such as data governance, which prevents potential problems from reaching end-users.

Improving Patient Outcomes through Better ETL

CASE STUDY

A Fortune 50 healthcare network with more than 150,000 employees delivers medical care to millions of members. It works with complex supply chains across a diverse array of medical specialties. It needs high productivity and responsiveness to provide high levels of care across a fast-changing medical landscape.

This leader in life sciences had gradually developed a tremendously complex IT infrastructure with more than 30 enterprise resource planning systems (ERPs), each configured differently. This complexity reduced data engineering productivity and caused delays in the many supply chains vital to delivering quality care.

This varied infrastructure required a series of individualized solutions to problems. The organization needed a simpler, standardized system that would democratize pipeline creation and data delivery and enable members of the data team. The organization chose Prophecy as its data engineering platform for its Databricks lakehouse. This solution has led to large improvements in productivity, with a single team now able to perform at the same level as three teams had previously.

Additional positive results of adopting Prophecy include

- A 66 percent increase in data engineering productivity
- A 650 percent reduction in pipeline development costs
- A much faster and smoother path to better investments

Members of the data team now use Prophecy to create flexible, scalable, and reusable data assets that support improved supply chain efficiency. This includes better inventory management, more efficient generation of sales orders, and improvements in manufacturing operations.

Finding MVPs in Major League Baseball

CASE STUDY

The Texas Rangers are an American League baseball team based in the Dallas-Fort Worth metropolitan area. The team has more than 250 employees. In 2001, the Moneyball era in major league baseball began, and today there's a commitment to player analytics across the sport. Teams need data for player acquisition and development as well as for between-game and even in-game decision making.

TECHNICAL STUFF

Moneyball is based on a statistical approach called *sabermetrics*, named for the Society for Advanced Baseball Research (SABR), leaders in player and game statistics. Statistics have even been used to drive recent changes to the game, such as the use of a pitch clock to speed up games.

The Texas Rangers went through several approaches to delivering and analyzing the statistics that teams need to thrive in the modern era. It began with a legacy, on-premises architecture. This setup was impossible to scale. As a next step, the team adopted a cloud data warehouse. However, this approach generated huge costs, delivered stale data, and required extensive maintenance efforts. Both approaches required complex data pipelines that could take up to two months to create, leaving the team without the answers they needed throughout a 162-game baseball season and beyond.

The Rangers moved to a Databricks Lakehouse architecture and adopted Prophecy. Now the team quickly ingests greater volumes of player data and rapidly extracts the insights it needs. This gives the club competitive advantage through greater productivity at scale.

Results of moving to a Databricks Lakehouse and adding Prophecy to the team include

- A 76 percent faster pipeline development
- A 10-fold data-ingestion increase
- Easy integration with standard services such as Git

IN THIS CHAPTER

- Designing and testing a pipeline
- Getting additional resources
- Watching a demo
- Starting your free trial

Chapter 6
Ten Resources for Getting Started

As adoption of the data lakehouse proceeds, organizations need to empower data users with the ability to perform data engineering tasks. Prophecy is a low-code solution that meets that challenge head on. To get a feel for what Prophecy can do for your organization and to get started, check out the ten resources in this chapter.

Explore Prophecy for Databricks

Prophecy has a video that shows you how to start using Prophecy within the Databricks environment. You can run pipelines directly on your Databricks Lakehouse and use Prophecy's visual canvas to build new pipelines and orchestrate them.

For more information visit `www.prophecy.io/prophecy-university?video=tibuwagqbn`.

Design a Pipeline

If you need help designing a pipeline, look no further. Prophecy's video gives you an in-depth guided tour showing you how to build a pipeline from a data source, including how to uses and configure Prophecy Gems.

To view the video, visit `www.prophecy.io/prophecy-university?video=08cek0x0nc`.

Test a Pipeline

Do you need to know how to test a pipeline you've built in Prophecy? You've come to the right place. Check out the following video to get started: `www.prophecy.io/prophecy-university?video=hs4r7qlsxo`.

Track Data Lineage

Tracking the lineage of data that flows through the pipelines you build in Prophecy is important. This step is important because it helps to ensure data quality and meet regulatory compliance requirements. For help in this process, you can view the following video: `www.prophecy.io/prophecy-university?video=etvhm3xvn8`.

Prophecy Documentation

If you like to read and want to know more in-depth facts about Prophecy, you can check out Prophecy's documentation. It provides a thorough description of the product, including differences between the Standard and Enterprise versions; concepts; and low-code approaches to Apache Spark and SQL. Access the documentation here: `docs.prophecy.io`.

Data Mesh Whitepaper

The data mesh is a distributed approach to storing and delivering data. The underlying platform may be owned by the organization's platform team or by domain-specific teams. In either case, the domain team has responsibility for its own data pipelines.

To help organizations implement a data mesh, Prophecy has created a guide that describes a practical approach to implementing the data mesh as a self-serve platform. Data teams can then use Prophecy to create their own pipelines that access and process data from the mesh.

Access the guide here: `landing.prophecy.io/implement-data-mesh-self-serve-whitepaper`.

Lakehouse Architecture Guide

This architecture guide shows how Databricks and Prophecy work together to provide you drag-and-drop access to data, built-in capabilities for data transformation, and much more.

Access the guide here: `docs.prophecy.io`.

Blog Post on Data Engineering

In this blog post, David Petrie at the Eckerson Group describes the modern architecture of the data lakehouse and recommends four principles for using it effectively.

You can access the blog post here at `www.prophecy.io/blog/data-engineering-for-the-data-lakehouse-four-guiding-principles`.

Request a Demo

If you think Prophecy running on a Databricks data lakehouse may be the right solution for you, the next step is to request a demo. Contact Prophecy at `www.prophecy.io/contact-us` and request your demo today!

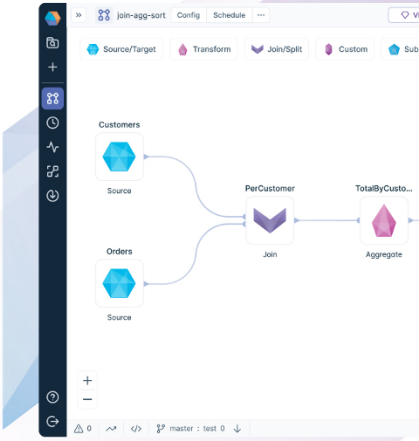
Start a 14-Day Trial

If you're already using a lakehouse, you may want to try Prophecy for yourself. Start your free 14-day trial here by visiting `app.prophecy.io/metadata/auth/signup`.

Prophecy

Low-code data transformation

Enable all data users to transform raw data into reliable, analytics-ready data using visual pipelines.



- **Low-code for all** - Empower every data user in your business to transform data like expert data engineers.
- **Trusted data** - Enjoy visual development that's based on software best practices and ensures data quality control for all data users.
- **Open and extensible** - Turn your data pipelines into open-source Spark or SQL code where you can add new visual components to standardize your operations.

Develop

Build low-code data pipelines that generate clean Spark or SQL code

Manage

Gain control with a metadata catalog, data quality, search and data lineage

Deploy

Reliably move to production with versions, tests, CI, CD and scheduling

Simplifying data transformation for all data teams.

Visit us at **prophecy.io** to start a free trial today

Data transformation for everyone

This book provides a primer into the power of low-code data engineering and how it can enable both technical and business data users with visual data transformation to convert raw data into analytics and machine learning ready data. You discover how data platforms like Prophecy are bringing this promise to life — make it simpler and faster for all organizations to democratize data access to meet the analytics and machine learning needs of the business.

Inside...

- Learn how data engineering has evolved
- Understand the value of low-code tools
- Learn about the data lakehouse
- Uncover top data engineering blockers
- Learn about Prophecy's solution
- Investigate customer applications
- How to get started with Prophecy

Floyd Earl Smith has worked in marketing at Apple, HSBC, NGINX, Onehouse, and Visa. He received his BA at the University of San Francisco and his MSc at the London School of Economics. He has written many *For Dummies* books, including *Quantum Computing For Dummies*.

ISBN: 978-1-394-20592-9
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.