LEARNING MADE EASY

Cloudera Special Edition

Analytics & ML in the Cloud

for dummies®
A Wiley Brand

Create a data catalog

Maintain a data flow that delivers new data

Manage the scalability of your data store

Brought to you by

CLOUDERA

Brett D. McLaughlin

About Cloudera

At Cloudera, we believe that data can make what is impossible today, possible tomorrow. We empower people to transform complex data into clear and actionable insights. Cloudera delivers an enterprise data cloud for any data, anywhere, from the Edge to AI. Powered by the relentless innovation of the open source community, Cloudera advances digital transformation for the world's largest enterprises. Learn more at www.cloudera.com.

Analytics & ML in the Cloud

Cloudera Special Edition

by Brett D. McLaughlin

for dummies®
A Wiley Brand

Analytics & ML in the Cloud For Dummies®, Cloudera Special Edition

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

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: WHILE THE PUBLISHER AND AUTHORS HAVE USED THEIR BEST EFFORTS IN PREPARING THIS WORK, THEY MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES REPRESENTATIVES, WRITTEN SALES MATERIALS OR PROMOTIONAL STATEMENTS FOR THIS WORK. THE FACT THAT AN ORGANIZATION, WEBSITE, OR PRODUCT IS REFERRED TO IN THIS WORK AS A CITATION AND/OR POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE PUBLISHER AND AUTHORS ENDORSE THE INFORMATION OR SERVICES THE ORGANIZATION, WEBSITE, OR PRODUCT MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING PROFESSIONAL SERVICES. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR YOUR SITUATION. YOU SHOULD CONSULT WITH A SPECIALIST WHERE APPROPRIATE. FURTHER, READERS SHOULD BE AWARE THAT WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ. NEITHER THE PUBLISHER NOR AUTHORS SHALL BE LIABLE FOR ANY LOSS OF PROFIT OR ANY OTHER COMMERCIAL DAMAGES, INCLUDING BUT NOT LIMITED TO SPECIAL, INCIDENTAL, CONSEQUENTIAL, OR OTHER DAMAGES.

ISBN 978-1-119-83533-2 (pbk); ISBN 978-1-119-83534-9 (ebk)

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.

Publisher's Acknowledgments

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

Development Editor:
Rebecca Senninger

Editorial Manager: Rev Mengle

Senior Account Manager:
Molly Daugherty

Production Editor:
Saikarthick Kumarasamy

Table of Contents

Introduction

Data analytics and machine learning. These terms are associated with the leading edge of technology, and in extreme cases, the basis for some of the most fantastical science fiction movies ever made.

Of course, like almost anything complex, when you break that complexity down into small discrete pieces, much of the confusion fades away. Data science and analytics is a deep discipline, but based upon the basic assumption that by grouping data and then looking at that data with specific tools and best practices, powerful conclusions can be drawn. These insights can range from how a user might use a web application to the tendency of a certain audience to pay for products on a certain day of the month.

In *Analytics & ML in the Cloud For Dummies*, you find out how to create a data catalog that can be used consistently, a data flow that keeps delivering new data to your catalog, and manage the scale of what can quickly become a massive data store.

About This Book

This book is about data analytics and machine learning. It's very direct and meant to be instructional. While you'll likely find you'll want other resources to go deeper into the various topics covered, this book can be your roadmap to the world of analytics and ML, easy to pull out of your back pocket or reference on your tablet at anytime.

You should consider this book as a jumping off point. Each chapter, and even each section, should cause you to generate questions and notes that will lead you into potentially deeper exploration of the topic at hand. Follow those questions and interests, and then come back to the next chapter, the next section. By the time you're finished, you'll have a clear understanding of how data analytics and ML can be used in today's modern, cloud-based businesses.

Foolish Assumptions

You may have heard what assuming does to you and me, but despite the old warning, I still make a few assumptions in this book.

I assume you have a basic understanding of the cloud.

I also assume you're an IT executive, a chief technology officer (CTO), a chief information officer (CIO), or another technology leader interested in organizing a framework for integrating data analytics into your systems and applications.

Icons Used in This Book

Throughout the book, I occasionally use special icons to call attention to important information. Here's what to expect:

REMEMBER

This icon points out important information you should commit to your nonvolatile memory, your gray matter, or your noggin — along with anniversaries and birthdays!

TIP

Tips are appreciated, never expected — and I sure hope you'll appreciate those useful nuggets of information.

WARNING

These alerts point out the stuff your mother warned you about (well, probably not), but they do offer practical advice to help you avoid potentially costly or frustrating mistakes.

TECHNICAL STUFF

This icon marks the places where the information gets more technical than you really need to know. Read if you're a geek and like the extra info; otherwise, freely skip without harm.

Where to Go From Here

There's only so much I can cover in 32 short pages, so if you find yourself at the end of this book, thinking, "Gosh, this was an amazing book! Where can I learn more?," check out `https://www.cloudera.com`. You can also find a great set of free and helpful training materials at `https://www.cloudera.com/about/training.html`.

IN THIS CHAPTER

- **Getting your data all into one place**
- **Unifying your data and your evolving deployments into the cloud**
- **Choosing tools that accelerate and enable your data's accessibility**

Chapter 1 Understanding the Data Lifecycle

In this chapter, you take the first steps to taking advantage of cloud analytics by involving data in your thinking as early as possible. You unify your data — because one dataset is almost always better for analytics than hundreds — and see how tools developed by others can be your best bet to move quickly and provide space for innovation.

Aggregating Multiple Datasets

The first thing you should think about when addressing your data and analytics needs is that of data source: Where is your data stored? What is the data source for your analytic tools?

Your data will never be in one single place

In any organization that values data, keeping your data consolidated is nearly impossible. You likely have cloud-hosted applications dropping data into an object storage service, local applications storing data in server farms, and logs on a variety of machines both local and virtual.

This leaves you with a few choices:

- **Combine your data.** Try and take all your data and copy it into a single data store. That's likely in the cloud and close to your analytics tools.
- **Adjust your data-creating processes.** Try to get all your applications to store data in a single location.
- **Create triggered copies of your data.** You could also look at building custom triggers that take data when inserted into one table (for example), and copy that data into a central location.

WARNING

None of these options are ideal, because all involve multiple copies of data, or changing the normal flow of your business and applications. This process becomes even more complicated when you have data both in a cloud as well as on-premises, or have intentionally set up a hybrid cloud.

Your data shouldn't have to be in one single place

The problem with all the solutions for combining your data isn't just the mechanics of *how* you'd combine the data. It's the underlying assumption itself. *Why* do you have to combine all your data? Often, the limitation is the tool or application you're using to consume that data.

HYBRID CLOUDS ARE NOT COMPLICATED

Ten years ago, the idea of a hybrid cloud was frowned upon. Organizations used hybrid clouds largely as a way to avoid fully committing to what was then a very new concept: cloud computing. The result was complex deployments and messy architecture.

Today, hybrid clouds are great solutions to combining the security of local data with the flexibility and computing power of the cloud. They also are much simpler and better architectures, and have lost that negative stigma.

Thankfully, you no longer need to put all your data into a single physical or virtual location in today's analytics world. You can instead build a *data catalog*, which is essentially a means of both identifying the type and location of important data, and a set of connectors to your data.

Your applications and analytics then interact with the data catalog, which — depending on the tool — can often also provide routing and fetching of data. You ask the catalog for any piece of data needed, the catalog retrieves the location of that data, and then facilitates the retrieval of that data. You get convenience *and* flexibility to store your data where you want, whether that's in a cloud or on-premises.

Connecting Local Data to Cloud Environments

It's relatively trivial to connect data in the cloud to cloud-based catalogs and processing tools. However, most organizations will not find it as simple to make local, on-premises data available to cloud-based tools.

Some data platforms and tools leave much of the work of getting data into their tools to you. This may mean you have recurring processes that copy or stream your local data to the cloud, or use storage that is backed by cloud storage and allow replication to handle the movement of data into the cloud.

Ideally, though, your data platform has components that can be installed on your local hardware. With tools that are co-located in your environment both locally and in the cloud, the integration of your data across multiple sources goes smoother. This also typically makes management of data easier, as common interfaces exist for all your data.

Streaming Data into Analytics Tools

Understanding the power of a data catalog to unify data that exists both on-premises and in the cloud is only one step in beginning a useful data lifecycle. There's another second and equally important step: You have to get data into your catalog and to your tools *fast*. With even five or ten applications contributing to your overall data store, your catalog is going to grow incredibly quickly. So how do you keep it current?

A data catalog that is stale is almost useless. But when you have thousands of new records every minute, even every second, how can you possibly keep it current? This is where you need a powerful technology partner. You need to stream metadata from multiple sources to your data catalog, and then stream actual data to your tools when your catalog asks for it.

TIP

Apache Kafka is a phenomenal solution here, so consider asking your cloud partner if they use Kafka to aid in their own streaming efforts. If you can't get your most current data to your analytics tools, then your analytics will never be real-time and that significantly limits their use to your business.

Innovating with Open Source Software

Nowhere is the power of people more important than in software. Development and architecture costs are always a significant part of any company's budget, and resources are nearly always constrained. Using open source software — software developed by an online community — works around these constraints.

By using an open-source project, your company and applications gain the benefit of that community, which often has skill sets and time that you do not. For instance, in the data world, the Apache Hive platform has been under development for years and provides a platform for distributed processing of enormous data sets and it's entirely free.

TIP

Through the innovation of a community, you can quickly integrate a framework such as Apache Hive and save thousands of hours of your own development time.

IN THIS CHAPTER

- Ensuring your data is always current
- Moving data from your applications into a data pipeline
- Connecting your data to the tools that analyze that data

Chapter 2 Building Data Pipelines

A *data pipeline* is a data store or dataset and the processing that occurs on that data store. Data pipelines are the core of data analytics; the key components that drive all your business insights. This chapter gets into more detail about building your own data pipelines as well as making them usable.

Moving from the Recent Past to Right Now

In the simplest scenario, you have a dataset and an analytics tool co-located. The data comes in — from an application or another processing tool — and is immediately consumed and analyzed. Your analytics are effectively real-time.

REMEMBER

Unfortunately, the world isn't full of simple scenarios. More often, you have to move data into a data catalog, potentially transform or filter that data, and then it's consumed on demand. All while more data is being produced, cataloged, and filtered.

Historical data is in the past

The problem with the more realistic multi-step scenario outlined is that it can result in stale data in a variety of forms:

- **Old data:** You could be analyzing data that isn't current because more current data hasn't yet been cataloged.
- **Unfiltered data:** You could be analyzing data that would be filtered out normally, but hasn't been to try and keep your data catalog current.

WARNING

That isn't to say that you want to ignore older or unfiltered data. However, you need to *know* that data is not current, and use it as a history; for example, to compare to current data to visualize a time series or predict user trends.

Current data requires continuous streaming

Chapter 1 shows how streaming data to your analytics tool is an essential part of a useful analytics platform. When considering the age of data, though — the "right now-ness" of data — that streaming has to be continuous. As one stream carries data from an on-premises disk to a cloud store, another stream may be processing data and transferring it to your visualization engine. Yet another stream could be adding new data to a machine learning process forming new models.

TIP

All of these streams need to be happening simultaneously. This is the only way to both keep historical data and differentiate that from the most current data.

Capturing Data at the Edge

Gone are the days when all your data was sourced from application logs or system writes. In the modern connected world, the Internet of Things (IoT) is constantly producing data. Smartphones, homes wired for automation, even refrigerators are all pushing data into the cloud.

DATA ON THE EDGE IS GROWING EXPONENTIALLY

Don't make the mistake of thinking that the sorts of devices collecting data at the edge are of little interest to your business. Ultimately, every business is about people. Whether it's those people's purchasing patterns or shipping them packages or facilitating business-to-business interactions.

Every day, more devices come online and feed data into systems that are listening. Be ready to collect that data at the edge when it becomes available!

Data at the edge is messy

Data at the edge (or sometimes called edge data) is generally unstructured, voluminous, and full of information that may not even be useful to your business. However, within all of that are potentially powerful insights that you will want to collect.

By using data pipelines, you can add steps to take this data and "clean" it. You can filter out information that isn't relevant to your business, transform and reformat it, and ultimately stream it into your data catalog alongside that less messy data from your own applications and systems.

Collecting data at the edge requires effort

This may seem an obvious statement: of course, collecting data from anywhere requires effort. But data at the edge is messy, and so the real question is, "*Who* exerts the effort to deal with that messy data?" Companies that are aimed at efficiency often rely on platforms and systems to incorporate edge data.

TIP

A good analytics platform provides listeners or other software devices that can collect heterogeneous data, transform it, and ensure it's ready for processing. In other words, let your platform partners exert their efforts so you can focus on interpreting the insights gained.

Managing Inward Data Flow

It's moving your data from the edge, into your data catalog, and then into your processing platform and tools that both shows off and requires a solid data pipeline. You should consider a data pipeline to be as much about the individual steps in that pipeline as it is about moving data.

While data pipelines follow a general pattern — data into data catalog, then fed into an analytics tool — they are also very flexible. Think of a pipeline as a set of steps. Some steps exist in almost all pipelines, but you can add additional custom steps. Perhaps you need to transform your data for processing, clean or filter the data, and then send it to a processor. All this is possible with a good configurable data pipeline, similar to the one shown in Figure 2-1.

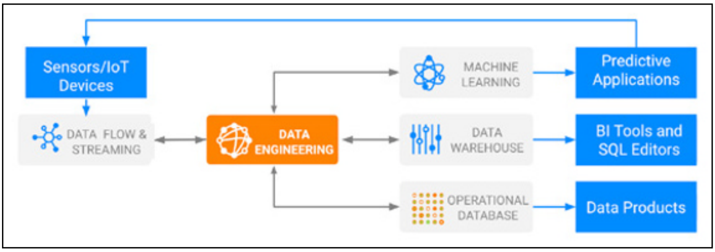

FIGURE 2-1: Good data pipelines allow you to insert multiple steps and configure each step to perform a job of data transformation or filtering.

As data moves inward from the edge, your job is to ensure that it ultimately looks like data from anywhere else in your system. Through data listeners and your platform choices to configurable pipelines, seek to create a catalog that makes all of your data available in a uniform fashion for analytics processing.

Feeding Data into Analytics Processing

At the end of most of your data pipelines you have your actual analytics processing. If you've set up your pipelines correctly, this step is the easiest: You have filtered, cleaned, structured data ready to be consumed and processed and ultimately turned into actionable insights.

However, even at this end of the pipeline, there could be additional processing steps: You may choose to store your data in a more neutral format and then apply a transformation suitable for an analytics tool to consume that data. This is particularly common if you're using multiple tools that consume data in different formats.

TIP

With a good pipeline, you can easily make these adjustments with another stage: Add a transformation. In some cases, the analytics tool can itself perform this transformation.

By ensuring your data pipeline is moving data into a catalog and keeping it consistently structured and formatting, any data — from your own systems, from applications, or from the edge and IoT devices — can all be treated and processed uniformly.

IN THIS CHAPTER

- » Selecting a great partner to accelerate your processing
- » Focusing on the right things in scaling your processing
- » Routing output data back into your data catalog
- » Visualizing results to find key insights

Chapter 3 Turning Insights into Action

The entire point of data analytics and machine learning is to improve decision-making. That's it! No matter how cohesive your architecture, no matter how clever your data pipelines, no matter how many devices at the edge are feeding your data catalog, the goal is better decision-making.

This chapter focuses on moving beyond the mechanics of data analytics and into the realm of application. You'll see that over and over, you have to keep a focus on actionable insights: outputs from your processes that make a difference in how you and your organization behave, and ultimately succeed.

Partnering for Machine Learning Success

The idea of a goal beyond the technical is perhaps nowhere more important than in making decisions about your analytics and machine learning (ML) approach. At a high-level, there are three different approaches you can take:

- **Do it all yourself.** This is for the hardcore engineers and analysts. You train your own models, tweak your processing and data pipelines often, and ultimately take complete ownership of your analytics and AI/ML operations.
- **Use pre-trained models.** A number of AutoML platforms provide pre-trained models that you can drop right into your process, or use these models as the basis for your own custom models.
- **A partner for processing and controlling your data.** In a more hybrid approach, you let a partner handle the configuration and low-level setup, and then you manage the modeling and tweak the outputs to your organization's needs.
- **Outsource everything.** In this approach, you essentially have a partner or data firm take your data and produce insights that are delivered "ready to act" back to you.

TIP

While there are some good arguments for the two more extreme options — doing it all yourself or outsourcing everything — you'll almost always gain the most benefit from a hybrid approach, where you and a partner work together to build systems and output usable insights.

AI and ML are hard to get right

The reality is that artificial intelligence is *not* an easy field. It requires expertise, experience, and the right tools. Not all organizations have all of this; in fact, *most* don't. Rather than try and solve every AI and ML problem you have, consider using a partner to provide the frameworks and tools, and in some cases, additional data expertise.

By removing the risk and cost of building a homegrown data analytics framework, you empower your own data engineers and analysts to spend their time on your business problems and delivering actual data insights. The return on this approach is often several multiples over what you spend on the partner who enables this approach.

All partners are not equal

When you decide to bring a partner into your project, you need to take into consideration these key items:

- **Experience in analytics:** This should be obvious, but it needs to be explicitly checked. There are a *lot* of options out there, and many potential partners are new to the analytics space. Look for a partner that has significant experience in data analytics.
- **Experience in your organization's industry:** Don't confuse subject matter experience with general analytics experience. Analyzing eCommerce purchasing patterns isn't the same as analyzing user access logs. Make sure your partner has worked with *your* industry before.
- **Experience with your cloud infrastructure:** If you're using a hybrid cloud and Google Cloud Platform, favor partners that can show they've worked with hybrid clouds on GCP in the past. There are differences between architecture and providers. Don't pay for a partner to experiment and learn on your dime.

TIP

It should be obvious here: experience is the key. But go beyond basic "we've done this a lot of times," and instead focus on the areas where a partner has worked with the specifics of your business before.

PARTNERS ARE TEAMMATES, NOT HIRED GUNS

REMEMBER

It's easy to build a pro/con list or a fancy decision matrix in Excel to evaluate partners. That's not a bad idea, but your decision shouldn't end with raw data. You want experience and a good partner that ticks your boxes for your needs, but you also should *like* your partner.

Just as a poor player can pollute a team's locker room, a partner that is knowledgeable but not bought into your organization's vision and purpose will almost always miss the mark. Ensure you evaluate your potential partners' work approach, team chemistry, and understanding of what you want out of your data analysis framework.

Almost every project starts with these as priorities. However, as timing and cost are provided, the initial priorities can take a backseat. Resist that! A project with the wrong partner can cost far more to undo and then redo with the *right* partner than a more expensive or less timely bid from the start.

Operating at Scale

By taking the partnering approach outlined in the previous section, you're likely already ahead in scaling your data infrastructure and ultimately, producing insights even when your data sources and catalog begin to grow.

Identifying key scaling requirements

These other scaling considerations are also critical:

- **Keep your data and processing secure.** Security is itself a large topic and is covered in Chapter 4. Suffice it to say that the security of your processes is a critical component in production analytics processing.
- **Avoid rogue tools and processes.** If you don't standardize your analytics tools, data engineers inevitably add the tools they need to get their jobs done and often from sources that aren't vetted and aren't built to operate at a large scale.
- **Insist on enterprise experience.** You've already seen all the ways in which experience can be critical to your organization's selection of a partner. When it comes to partners and tools, dealing with petabytes of data across an enterprise is far different than a few gigabytes in a one-application environment. Make sure your platform can handle both.

Using hybrid clouds to scale over time

The growth of hybrid clouds is a critical development in scaling your own environment and analytics. With well over half of all major organizations moving toward hybrid clouds, you can stagger your scaling needs (`https://blog.cloudera.com/why-adopt-a-hybrid-multi-cloud-strategy/`). Moving data to the cloud is both simple and an easy way to scale storage without adding giant server arrays.

At the same time, keeping critical and security-sensitive portions of your application running on-premises means you don't have to scale *everything* at once. Scale your data and analytics quickly by moving them to the cloud where processes are clear cut, and then take a longer approach to get your applications hosted where it makes the most sense.

TIP

A similar approach is to use multiple cloud environments. While this will require a lot more knowledge of different cloud architectures, you may find that one cloud — such as Google — is better suited for data analysis but another cloud — such as Amazon or Microsoft — provides services that make cloud hosting of one of your applications easy. Don't be afraid of these multi-cloud architectures.

Building consistent interfaces across environments

When you do embrace hybrid or multi-cloud architectures, you need to ensure that your data analysis tools and control interfaces are consistent, such as the dashboard shown in Figure 3-1. If you have a different dashboard and process for on-premises compared to Google, and then again compared to Microsoft, you're going to have a management nightmare and that will absolutely not scale.

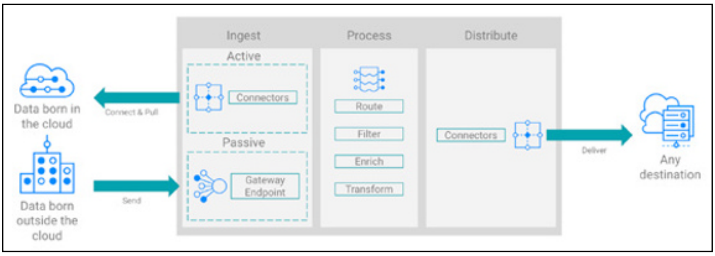

FIGURE 3-1: Manage your environments and data across clouds and on-premises using a single unified dashboard or console.

When you think about unifying your data into one data catalog, this is just an extension of that same principle: You want one interface (or at a minimum, consistent interfaces) for managing and analyzing that catalog.

Ensuring Insights Are Available Everywhere

Much of this chapter is about ensuring the data you have — in a variety of environments, on-premises and off — is available for processing. But it's one thing to unify your catalog and scale the processing of that data; it's something else altogether to ensure that there are no limitations on the consumption of the resulting analytics.

Analytics are more than visualizations

It's easy to think of the output of good analytics and AI as a dashboard full of visualizations. You can glean an idea of what types of interfaces users are most responsive to, the lifetime value of an average customer, or even what workflow is most efficient at driving a potential customer to make a contact request.

But while you can see those things on a visualization, the insights themselves are just additional data. That data should be used to drive decisions, both through seen visualizations and as input to your internal processes and external applications.

This means that the output data of your processing tools must be fed to other processes, and potentially back into your data catalog itself. It then becomes input data for even more refined analytics.

REMEMBER

OUTPUT DATA IS STILL DATA

Even insights about your data is still data and data belongs in your data catalog. This is an important learning. The best organizations are constantly growing their data catalog, often exponentially, through collecting raw data and also storing any insights on that raw data as second-level (or derived) data. The result is an even broader set of potential findings.

Keeping insights actionable

Don't think that storing output data in your catalog means that visualizations aren't critical. By examining analytics visually you can keep your focus on what changes and improves your business. You often need to change your calculations, collections, and metrics many times to arrive at the right level and types of insights.

TIP

This is where your entire data pipeline reveals its power. The more you can — ideally in real time — query your data catalog in different ways and see what pops out, the more you will find useful ways to use your data.

This is also why you want unified interfaces across your data stores, on-premises, and in multi-cloud environments. You need to "play" with findings until something stands out. Then, because you have ensured that your analytics results are also accessible as data across your platform, you can employ those findings in ways that move your business forward.

IN THIS CHAPTER

» **Building a secure data and application architecture**

» **Avoiding differences in architecture through replication and change control**

» **Building trust in public and hybrid clouds**

Chapter 4 Securing and Integrating Your Data

You've collected your data, made it accessible via a data catalog, and even performed processing on the data. You've cycled your results and insights back into your data catalog. So what's next? You need to take that data and push it to your applications to use, leveling up your organization in the process.

Along the way, security is an ever-present concern. Nowhere is this more true than of the data you've collected and the actionable decisions you're made through that data's use. If your data isn't secure at every stage of the process, you risk loss, noncompliance, and ultimately, a failure of your entire analytics platform.

Securing by Design

There are two basic ways to handle security:

- **Security by design:** In this approach, every bit of the architecture and infrastructure of your platform and application is built with security as a primary consideration.

- **Security by layering:** Also sometimes thought of as "security after the fact," this typically means that once infrastructure and applications are built, security is layered on top. It's not that security is unimportant, but that building the application takes precedence.

Inconsistency is the enemy of security

Security thrives on the known. The more that is known, the easier it is to secure (by design). When things are constantly changing or slightly different from another, security has to adjust, and that's usually not a good thing.

This is especially true in terms of architecture. You should look for and choose solutions (or, if you have to, build them yourself) that behave the same on-premises and in the cloud. The databases should be structured the same, clusters of servers should use similar options and ideally be managed by the same agents, and your infrastructure diagrams should be mirror images of each other.

This doesn't mean that your applications and analytics are cookie-cutter, but it does mean that the systems underlying them are. If Kubernetes manages your virtual servers, then Kubernetes should manage your on-premises servers (through virtualization). If you store your unstructured data in an object-oriented database (OODBMS) on-premises, then you should use an OODBMS in the cloud as well.

Favor replication over installation

One great way to ensure security by design is to replicate environments in different contexts. Rather than installing a server cluster on-premises and then repeating the process in the cloud, build an environment in one place and replicate it to the other contexts. This means that you're essentially "copying" an environment over and over and that copying ensures consistency.

The result is that the security policies built for one environment will apply to all your environments. The consistency here multiples the effect of your security work by ensuring it applies to all of your environments — even ones that haven't been built yet.

TIP

CHANGE ONCE (AND ONLY ONCE)

With consistent, replicated environments, the location of any changes itself must change. You should have a single test environment that you can tweak, and then take any successful updates back into your core platforms. Those updates can be applied to all your environments — again through replication — to ensure consistency.

You must avoid the temptation to make "one quick change" to any running environment, as that is exactly how security holes form. Instead, exercise the discipline to change your controlled test environment, verify things work, and then replicate that change across all your environments with confidence.

Track and audit everything

In addition to designing a system where changes are centralized and applied uniformly, you need to track and audit everything that occurs in your system.

REMEMBER

Consistency of design doesn't replace the need for audit trails.

When you centralize your data catalog and make your environments consistent, your system can begin to keep up with data lineage, more metadata about data classification and usage, and data metrics. These all form data-centric audit trails that explain where data came from, how it got to its current state, and how it's being used. When that data is processed and analyzed, resulting insights can further be classified, including the data from which those insights were derived.

Overcoming Fears of the Public or Hybrid Cloud

If you find yourself in a large or more technically conservative organization, or one that already has recognized serious security concerns, you're likely to run into some resistance when it comes to employing the public cloud.

REMEMBER

If you’re moving to a hybrid cloud, then all these same objections apply to the portion of your target architecture that *isn’t* on-premises. In other words, don’t skip this just because you’re not completely aiming for a public cloud architecture.

Common objections to the public cloud

Most of the objections to the cloud fall into one of a few categories:

- **The public cloud is insecure.** Public clouds are often plagued by the term “public.” “Public” is often equated with “accessible by everyone.” That doesn’t sound like a recipe for security, especially when it comes to data.
- **The public cloud is not controllable.** This is a tough one. On-premises data and applications mean the control of those systems is in-house. No stranger is in a dark room somewhere handling upgrades or architectural decisions.
- **The public cloud is different.** With managed services and virtual hardware and different access patterns, the public cloud isn’t just a virtual closet. It’s a new way of working, and that means education, mistakes, learning, and ultimately, time and money.
- **The public cloud is an unknown.** This is a bit of a catchall, but you’ll find it over and over. The cloud is still not well-understood by many non-technical leaders, and no technical explanation can overcome uncertainty. Because it supposedly isn’t secure, well-controlled, and similar to what currently exists, then it must be a risk, and many leaders are deeply risk-averse.

Turning objections into best practice

You’ll have to combat these objections, but thankfully, many of these battles have already been won by others over and over.

- **You still secure your data in the cloud.** Public clouds secure the environment, but you can still secure your own data (and applications) to the degree you prefer. Data encrypted at rest and in transit, private keys, access management — you can do everything in the cloud that you can do locally.

- **Your cloud is controllable by you.** The public cloud is available to the public; your applications and data hosted within it are not. This critical distinction must be made repeatedly. Even better, you can control access to your area in the cloud far better than a server closet in a physical office.
- **The cloud is modern (and different).** Be careful to not argue that the cloud is not different, because that's not true. Argue instead that the modernity of the cloud means you'll be getting access to cutting-edge technology and the results of many other people (who you didn't have to pay) working to advance data security and analytics.
- **Your own systems were unknown once.** When it comes to the risks you *don't* know about in the cloud, that's no different than any system. Affirm that there are unknowns everywhere, but a hybrid cloud, in particular, allows you to mitigate those risks with on-premises systems until a greater level of comfort is reached.

Prioritizing a Single Data Context

In the same way that you need to ensure a consistent data architecture across environments, the same is true for your data permissions and structure. Think of this as your *data context*: everything about the security and governance of your data, including metadata and permissions.

This context then is central to all of your environments, and is shared across all of them. It's essential to have a single data context to ensure consistency of your data and your systems in their entirety.

Consistency is critical

The consistency of a single data context provides you a number of critical components:

- **A robust data catalog:** Your data catalog is central to all your analytics. Your data catalog, then, is the table of contents, index, and cover for that catalog. It ensures your catalog is secure and the data within it discoverable.

- **A secure catalog and environments:** Data access across environments involves a *lot* of permissioning. A centralized, singular data context gives you both granular access control and ensures that access is uniform for all your environments.
- **Data governance:** Governance goes beyond raw security to provide regulatory compliance. Your data context adds classification and audit trails for your catalog.
- **Consistent and complete data encryption:** While clouds often provide security, a data context can ensure consistency of security in all environments, cloud or on-premises. This should apply to data at rest and in transit.
- **Unification across hybrid cloud:** Any gaps in your architecture due to on-premises variances from cloud can be handled by a singular data context.

Advantages of aligning data across architectures

When you use a data context in this manner, you get more than the qualitative advantages described in the previous section. You also should see a lowering of operational costs and a hardening of your system's security posture.

REMEMBER

Consistent environments always result in reduced costs, as well as making security easier through consistency.

You will also see improvements through centralized management of your systems. They're all the same, and a change to one can easily be deployed to all. You should also ultimately gain better data insights, because results won't be skewed by odd variances between data or overlooked data points because of a non-shared data catalog.

Moving from Data to Applications

Ultimately, the insights you gain from good analytics can be actualized in two ways:

- **Through visual inspection and out-of-code decisions:** You can use dashboards and visualizations to examine insights and make organizational decisions, such as prioritizing specific features in applications, avoiding costs in less profitable segments, or marketing to specific customer bases.

- **Through consumption by existing and new applications:** You can take output data and funnel it into applications that in turn make automated decisions based on that data. Applications respond in a more tuned fashion to requests and become more responsive and accurate based on collected data.

Embracing transformation to effect improvement

While the first method of using data insights — viewing them through visualizations and dashboards — is intuitive, it is only a first step. It also requires manual inspection of results and accompanying business decisions, all of which take time and oversight.

The second approach, where insights are pushed into your applications for immediate use, is far more powerful. However, this approach likely causes a greater need for application and business transformation.

You need to potentially refactor your applications to take these insights as input data, and then use the insights in meaningful ways. You also have to ensure that all your applications are more data-driven and less static than is typical.

REMEMBER

If your insights don't cause an actual change in how your applications function, then the insights aren't truly creating action in your organization.

Taking advantage of data lakes to unify data access

One relatively easy change you can make is to ensure that all of your data — whether it's collected and generated, input or output — is accessible through a data lake that provides uniform access. If your applications can access data that they already are consuming along with new data through uniform interfaces, it's much easier to make use of the new data.

TECHNICAL STUFF

A data lake is similar to a data catalog in this context: The catalog makes discovering data simple and possible through a single interface and approach, and the data lake makes actual access to that data equally consistent. For an application, consuming user input data, log data, or insights gained from that input and log data should follow the same exact process.

Adding self-service data to your platform

Another key component of ensuring your data is actionable at both the manual and application levels is enabling self-service. Through your data catalog, automatic permissioning, and governance through a data context, you can make your insights available via self-service. This means that applications can pull data in an ad-hoc fashion rather than only having access to predetermined data sets.

This also requires everything that has been already discussed working in concert: a primary data context, a centralized data catalog, a data lake, and data governance at the context level. This has the advantage of forcing any data silos to be broken down and made discoverable across your organization.

IN THIS CHAPTER

- » Keeping your architectures, environments, and data consistent
- » Ensuring your infrastructure and data catalog are current and well-organized for lower costs and faster performance
- » Making effective use of cloud-native services and functionality

Chapter 5 Ten Analytics and ML Challenges

Here are ten challenges you'll face and overcome to build a successful and actionable analytics and ML platform.

Dealing with Multiple Architectures

You've seen this issue come up in nearly every chapter and context. Hybrid clouds and multi-cloud environments are now common and must be accounted for. You should always have consistency in your deployments and architecture and centralized management of your environments, whether on-premises or in the cloud. Additionally, you should make changes centrally and deploy those changes everywhere, because consistent environments won't *stay* consistent unless you make a point to keep them that way.

Keeping Datasets in Sync

When you have data across multiple architectures and environments, it's possible to have discrepancies in that data. First, prioritize a single data context and data catalog and manage both in a single place. Then, favor referencing data through the catalog rather than making copies of it in different locations. The more you rely on a single source of data (with backups being kept for recovery rather than normal use), the fewer inconsistencies you'll have in your data.

Storing More Data for Analysis

If you've primarily had on-premises architectures, you've at some point likely had to make choices about what data is stored and what is discarded. This relates to cost. Adding hard drives, keeping them powered, and maintaining space for enclosures is a significant OpEx cost. However, in the cloud, the cost for additional storage is low, and you can take those savings and apply them to services that supplement your handling and analysis of that data.

TIP

You should generally opt to *keep everything* and keep it readily accessible (or "hot") for at least 30 days. You may see storage cost savings by moving older data to cold storage over time, but in general, you can store everything for fractional costs, and more data means more opportunities for insights.

Avoiding Modeling Mistakes

As you analyze more data and bring analytics and machine learning to bear, have confidence in your processes and models. However, trust but verify. If a model and the findings it generates make sense, understand *why* it confirms your expected findings. If the findings are surprising, again, review the data and explore why your own assumptions were incorrect.

REMEMBER

Never take results at face value; in this way, you avoid outliers that skew your results or mistakes in your model.

Choosing a Partner

Much was said in Chapter 3 about key factors in choosing a partner to assist your analytics needs. It's worth adding to that: *cost is rarely the primary driver*. While there are times that the lowest option is the best — the bidder may be new, have uncovered a cost-saving innovation, or underpricing an initial bid in favor of a long-term relationship — that is the exception rather than the rule.

Try and determine your differentiators *before* you receive bids, and remove cost as a differentiator purely as an exercise. If you find that your key decision-making points are things like experience, segment knowledge, or architectural compatibility, you're likely on the right track. Resist the urge to favor fewer material issues such as location or price point.

Avoiding Unneeded Overhead

While the cost for storing data is tiny in a cloud-favoring world, you can still spend millions on cloud services and hosting. Data is cheap but architecture is not. You and your partner should constantly be examining your virtual footprint — particularly how many instances are being used by your analytical processes.

WARNING

As a general rule, modern data analytics platforms take advantage of streaming data and serverless technologies. If you're seeing hundreds of virtual instances, you may have an outdated architecture. And if your data is being copied to or staged in multiple environments, ask about streaming and don't settle for evasive answers.

Optimizing a Data Catalog

Your data catalog, along with your data context, is the heartbeat of your system. While a context is largely self-organizing (permissions and access levels are naturally easy to store and manage), your data catalog is not. Additionally, the constant influx of references to your data into a catalog means that the catalog is often busy, well, cataloging, and not organizing.

Any good platform has a means of keeping your catalog organized and ensuring its references and inputs are refreshed, and that stale references are purged. There are a number of ways to do this, but the point is that it needs to be done, and done often. Whatever the approach, an optimized and clean data catalog will dramatically out-perform one that isn't.

Taking Advantage of Managed Services

If you can use a cloud-native service instead of bare virtual servers and installed processes, you generally should. Whether it's a managed database or data warehouse or a code runner such as Amazon's Lambda rather than a self-hosted interpreter, you'll have less to manage and spend less by taking advantage of managed services.

These services don't have to be from your cloud provider, either. Your analytics partner should also be providing native services that are business-oriented rather than virtual hard drives and servers.

Choosing Between Hybrid and Full Cloud

This is a bit of a false heading; you really shouldn't *choose* between hybrid and full cloud. Rather, you should always be evaluating what your best mix of hybrid and full cloud is and keep evaluating. Often, you're in a constant state of moving resources into the cloud, or multiple cloud environments.

Don't be satisfied with a singular decision at one time. Instead, as new systems, tools, and applications come online, make the best choice with the new information you have. And in light of this, *insist* that your tools and platforms interact with hybrid, all-cloud, and multi-cloud environments.

Measuring and Quantifying Results

This seems almost silly to say in a book on analytics, but it needs to be said anyway: Measure everything! Don't just gather analytics for your business. Gather them on your operations. Track your costs of operation, the influx of new data, the growth in customers as well as the growth of your data store.

In every case, business and operational, you should be evaluating your measurements in light of tangible and concrete goals. If your results don't match up, it may be okay; but understand *why* there is a mismatch, and make intentional choices about whether to change what you're doing or continue. Run your business on numbers, and that should include your operations and analytics themselves.

I DISCOVERED
how to prevent maintenance problems from becoming passenger problems.

Get actionable insights into your data

Managing your organization's data is essential, but has become complex as data grows and is stored in multiple places. A platform that is cloud-agnostic, has a security and governance underpinning, and helps you define a data and analytics strategy is a real asset. Your organization can manage, provision, and enable powerful analytics. With *Analytics & ML in the Cloud For Dummies,* you'll have a clear understanding of how data analytics and ML can be used in your modern, cloud-based business.

Inside…

- Build a data catalog
- Manage data through its lifecycle
- Move data through a pipeline
- Use data to make business decisions
- Choose a hybrid solution for your data
- Scale up as your operation grows
- Keep your data secure

CLOUDERA

Brett McLaughlin has been in technology for 25 years. He spent 8 years leading NASA's efforts to move its massive earth-sensing satellite data into a modern cloud platform. He is an ecommerce veteran, serving as CTO for Volusion, then sticky.io, and now KlickTrack.io, transforming each organizations' technology and product organizations.

ISBN: 978-1-119-83533-2
Not For Resale

9 781119 835332

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT