Chapter 10

Big Data Stewardship and Data Lakes

INTRODUCTION

With the advent of “big data” and the move to implement “data lakes,” a natural question that comes up is how “big Data Governance” (or “big Data Stewardship”) is different from “regular” Data Governance or Data Stewardship. The assumption on the part of those asking these questions seems to be that there *must* be big differences, and that the roles, procedures, and metadata captured are also quite different. And, that these differences could require a wholesale restructuring of Data Governance and Data Stewardship efforts. But nothing could be further from the truth. While there *are some* differences (as will be discussed in this chapter) most of what has been built and implemented will remain largely the same. Big data and data lakes make Data Governance and Data Stewardship even more important because much of the promise of these technologies requires *more* Data Governance and Data Stewardship in order to be valuable.

THE FALLACY OF UNSTRUCTURED DATA

Another common misconception is that “unstructured data” somehow changes the importance or need for Data Governance and Data Stewardship. What is often meant by unstructured data is textual data that is written out by various processes and not stored in a rigid structure such as a relational database. For example, web logs often contain valuable information (which is recorded nowhere except in the web log) such as who has logged into the system, how often, and what web pages were accessed. But from the standpoint of Data Governance, this data is *not* unstructured! In fact, it has a well-defined structure, which is set by the program writing out the text. The process for changing that structure is quite different from a relational database (and some would say it is easier)—“simply” change the code that writes out the data. No need for new tables, new columns, changes in datatypes, or adjustments to foreign keys. But make no mistake—the data has structure and is largely worthless unless that structure is supplied via metadata to those who need to use the “unstructured data.”

(*Continued*)

Data Stewardship. DOI: https://doi.org/10.1016/B978-0-12-822132-7.00010-3

(CONTINUED)

And, just like any other kind of data, if the information recorded in this type of "unstructured data" is considered valuable enough to be governed, then all the same steps need to be taken to bring it under governance.

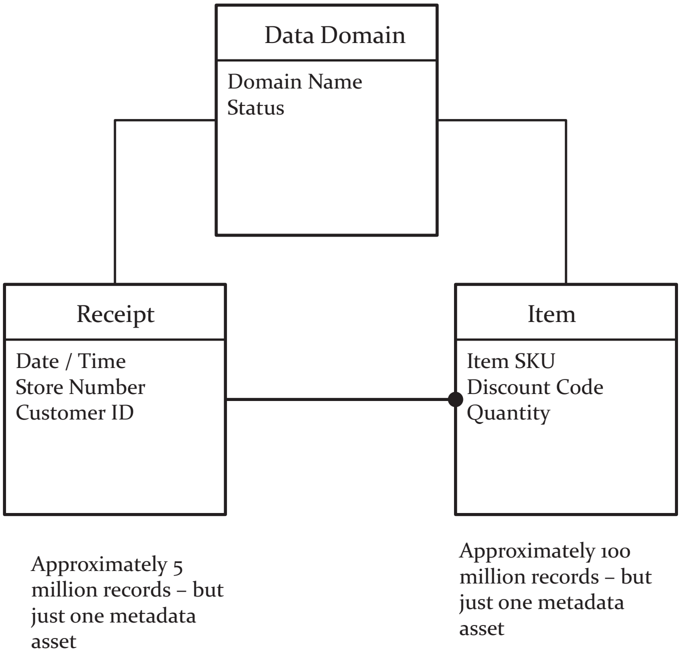
Another example of truly unstructured data is documents (such as a Word document or a PowerPoint presentation) or "objects" that are simply stored as is with the intent that they can be found later and opened in the originating application or something similar. But here again, when it comes to governance, metadata about the documents or objects is what is governed, and that metadata has structure. For example, in order to know the business function that would steward a Word document, something about the document must be known and recorded such as the title, topic, subject, and sensitivity of the document.

DATA STEWARDSHIP AND BIG DATA

Over the years, many companies have found it valuable to capture ever-increasing quantities of data. This is both because the infrastructure (networks, storage, and easily-changeable configurations for structures) has matured and gotten vastly less expensive, and more importantly, because companies have realized that a lot of value can be gained by analyzing and comparing the data collected. I say "more importantly" because if there was no value in collecting large quantities of data, they wouldn't do it, no matter how cheap the infrastructure becomes.

A good example that almost anyone can relate to is a chain of grocery stores. The register records every item that is purchased, how the purchases are grouped (onto a single receipt), and even (when the grocery chain has a loyalty card program) *who* purchased the items. As you can imagine, this leads to a vast quantity of highly detailed data from which the grocery chain can extract value. For example, by analyzing lists of items that are often purchased together (such as frozen waffles and syrup), the grocery chain can provide coupons that are both more useful to the customer and more profitable to the company. I (and probably you) have seen this sort of analysis *not* carried out properly. Thus I might buy a box of frozen waffles, and the store gives me a coupon for a free box. This makes no sense because I have proven that I'm willing to pay money for the waffles, so why give it to me for free? Instead, if the "big data" available shows that waffles and syrup are often bought together, and I buy waffles but *not* syrup, it would make more sense to give me a free (or discounted price) on syrup. This outcome helps to get me thinking of buying these items together (good for the store) and provides a discount to do so (good for me).

The point is that in the days when it was impractical to keep the purchase data at the level of granularity needed to perform such analysis, companies may well not have done so, settling instead on the summaries and aggregations needed only to control inventory and ordering. The business data elements and the rest of the metadata needed to govern and steward the data would not have included the granular data (such as individual retail transactions, grouping by receipt, and identification of the customer) that can add value today. But in today's world, where the value added by such granular data returns value to the company, the additional business data elements and other metadata *would* be collected, governed, and stewarded (as shown in the example in Fig. 10.1). Thus, once again, Data Governance and Data Stewardship are more important than ever before—though the processes and roles are largely the same. In addition, the rest of the processes that include Business Data Stewards (such as data quality improvement) are larger—but not really any different. As illustrated here, *more* data collected usually means more metadata collected as well.


Data Domain
Domain Name
Status
Receipt
Date / Time
Store Number
Customer ID
Item
Item SKU
Discount Code
Quantity
Approximately 5 million records – but just one metadata asset
Approximately 100 million records – but just one metadata asset

■ **FIGURE 10.1** More granular data requires a small additional amount of metadata, but possibly a *lot* of extra data.

DATA STEWARDSHIP AND DATA LAKES

With the advent of "big data" and the need to be nimbler when working with data came an architectural construct called a "data lake." It is a system that holds data that is collected from various source systems and all stored in that one system. In some ways a data lake is like a data warehouse, in that part of the data lake stores data that has been combined from other systems. But unlike a data warehouse, a data lake stores data in various zones, including "raw" (looking like it does in the source system) and various transformed forms. These transformed forms (which often combine data from multiple source systems) are specialized for tasks such as reporting, visualization, analytics, and machine learning.

NOTE

Whole books (many whole books) have been written about what a Data Lake is, the advantages it has over classic data warehouses, and how flexible and inexpensive (compared to a monolithic data warehouse) data lakes can be. But this is not a book about data lakes, except insofar as Data Stewardship plays an important role in data lake governance.

Data Lakes and Metadata

As you might expect, metadata plays an important part in the use and governance of a data lake. A data lake is a data management platform, and as with any other data management platform, the data within it must be described, understood, owned, and inventoried. Lineage must be understood to prove integrity, and in most of the zones, the data quality must be known. In fact, in many companies, data cannot be loaded into the production data lake (a process called "ingestion") until the metadata has been collected and documented. Many data lakes even incorporate a metadata "catalog" to hold all this required information.

NOTE

Isolating the metadata for the data lake within the data lake itself creates a silo of metadata, separate from the rest of the company's metadata. Although siloing your metadata is usually not a good idea, with some data lake architectures it may be unavoidable.

You may have also heard the term "data swamp," which is taken to mean a data lake in which the appropriate metadata management is not in place, and that lacks Data Governance. Much like a data warehouse, Data Stewardship and curation of the data is needed, along with metadata capture.

DEFINITION

Curation is the action or process of selecting, organizing, and looking after the items in a collection or exhibition.

Sound principles of data definition, data production, and data usage are also needed. This is because the usefulness of a data lake is driven by knowing certain things about the data sets in the data lake, including:

- What is available in the data lake? Like any other data store, you have to know what is available before it can be useful. Knowing what is available includes not only what the data means, but what systems served as sources for the datasets, how those datasets were transformed and combined, and the quality of the data in the dataset. Much of this information (metadata) is contained in the "catalog".
- How can I find the data I'm looking for? The same data may be available in multiple zones of the lake. As we'll see, zones are handled differently (different levels of data cleanliness and different

levels of Data Governance). Thus you not only need to know what data is available, but which zone it is in, or, if it is in multiple zones, which zone is most useful for your purposes.

- How was the data selected? Different data is selected from different sources, combined in different ways, and ingested into the data lake. Depending what how the data was selected (and where it was selected from), you may find certain datasets more useful (or less useful) than others.
- What access is available? Getting at the data in a data lake requires sophisticated processes that make the access available. There are different publishing methods and distribution systems as well as different levels of access. For example, you may not have access to certain repositories of data because they contain sensitive or private information that you do not have permission to see and use. You may be able to get access, but you'll have to go through the appropriate procedures to do so.

From another perspective, the same questions you would ask if you were doing analytics in a data warehouse also need to be answered if you're doing analytics in a data lake. These questions highlight the need for rigorous Data Stewardship, at least in any of the zones where analytics are being performed.

These questions include:

- *Who owns the data and can answer questions about it?* This is a classic question that users ask and one of the most obvious drivers for Data Stewardship. In the presence of a robust set of tools to record metadata, this information can be found in the Business Glossary (possibly in combination with a metadata repository).
- *How do I find the right data elements that meet my needs?* Once again, robust Data Stewardship (and a good Business Glossary/ metadata repository) provides the answer to this crucial question. The business user determines the data they need to solve a particular business need, and the results of Data Stewardship provide the location of the needed data.
- *How do I clean the data to an appropriate level of quality?* In consultation with the Business Data Stewards (either by business function or by data domain—as described in Chapter 11: Governing and Stewarding Your Data Using Data Domains) a determination can be made as to the level of data quality that will be sufficient to support the business need. After that determination, the actual data quality can be ascertained by data profiling, as described in

Chapter 7, The Important Roles of Data Stewards. If the data quality is insufficient to support the business need, an issue can be raised, and a project started to mitigate the issue and bring the quality up to the level needed.

- *What is the right security for the data being used?* The compliance and privacy parameters for data should have been established when the business data elements were defined. If these parameters were *not* set, then it will be necessary to consult with the subject matter experts (who are likely situated in your Risk, Privacy, and Compliance organizations) in conjunction with the responsible Business Data Stewards to establish those parameters during ingestion. The question then becomes whether the data analysts trying to use the data are allowed to have access to the sensitive data. If they do not have the necessary access, then the appropriate processes (as described in Chapter 7: The Important Roles of Data Stewards) would have to be followed to gain access to the data.
- *Is the data being monitored for adherence to standards?* The Data Governance/Data Stewardship effort is responsible for adhering to the standards and policies related to data. Usually, a different group will be responsible for monitoring whether the adherence is adequate and reporting on any shortfalls along with designing mitigating steps. Thus although monitoring for adherence to standards is usually done outside of the Data Stewardship effort, that adherence—and the advantages that come with it—starts with Data Stewardship.

Determining the Level of Data Governance for the Data Lake

As we've already mentioned, the metadata about the data in the data lake as well as the level of Data Governance for the zones are decisions that need to be made. Each zone may have a different amount of curation and "strictness" of Data Governance. To a certain extent, the metadata and level of Data Governance are something of a balancing act. Overdoing the amount of metadata collection (whether in the catalog or elsewhere) stifles the usefulness and flexibility of the data lake due to the overhead of collecting the metadata before bringing in a new dataset or modifying an existing dataset. Put another way, the more you know (metadata) the more useful the data is, but also the more work it is to add datasets and make changes to structures. Since ease of adding data and flexibility of structure are two of the primary advantages of using a data lake in the first place, you must balance the benefits against the cost.

What is likely to happen is that the metadata collected and level of Data Governance will be set based on the use of the data in the zone. The goal is to use the data with a level of confidence consistent with that use. Things to think about include:

- There is a need to tune the Data Governance to the priorities and context (use) of the data. For some people, a level that some would consider a "data swamp" might be perfectly acceptable for their needs.
- Certain decisions drive the data lake in one direction or another. For example, if a zone in the data lake becomes the source of data for regulatory reports, it needs extremely good metadata, cataloging, Data Governance, and lineage.
- In some zones, lightweight Data Governance on adding, naming, and curating the data protects the data lake shared resource from the "tragedy of the commons."

DEFINITION

The "tragedy of the commons" is a situation where individual users, acting independently according to their own self-interest, behave contrarily to the common good of all users by depleting or spoiling the shared resource through their collective action. In the case of a data lake, insisting on perfect metadata and rigorous Data Governance is inconsistent with the purpose of a data lake, and doing so would ruin the data lake for the users who don't need that and that can't support that effort.

Metadata Created in the Data Lake

As data is combined in inventive ways in the data lake to serve the business needs, the question arises as to who is responsible for the newly created/derived business data elements. This is especially true in an organization with a strong Data Governance operation since the idea that the business data elements should remain largely undefined and that no one is accountable for the data is repugnant and may even violate data management policies. The issue extends especially into data quality since this "new" data is often created for new purposes or business drivers, and there is no guarantee that the data quality is sufficient for these new purposes and, thus, must be profiled to find out. Closely aligned to the evaluation of the quality of the data are the data quality rules against which the data will be evaluated—and there must be an accountable business function or data domain to define those rules.

So who is responsible for the metadata that must be generated as data is manipulated within the data lake? Figuring that out can be problematic because by the time the data reaches the fully integrated zone, it includes data from many source systems and, thus, integrates the knowledge/metadata expertise of many different Business Data Stewards. This integration must be done carefully and well and linked back to a common vocabulary of business data elements. This common vocabulary must be enhanced as new business data elements are defined and derived.

The data is initially brought into the raw zone of the data lake from source systems in a process called “ingestion.” Where the source system(s) providing the ingested dataset do not have all the business data elements defined and governed, many companies will demand that it be done when brought into the data lake. This would include establishing the privacy and compliance parameters of the business data elements, updating the physical location as being in the raw zone of the data lake, and monitoring adherence to standards. In addition, when the data in the source system did not have data quality rules defined and used to profile that data, that may be done during ingestion as well.

NOTE

There now exists a whole class of tools that will combine ingestion into a data lake and data profiling as part of that ingestion.

When the data in various zones are combined and integrated together to form the contents of another zone, these same processes must also be carried out; though depending on the “rigor” of the required Data Governance in the zone, some steps may be skipped.

Fig. 10.2 illustrates how this might look.

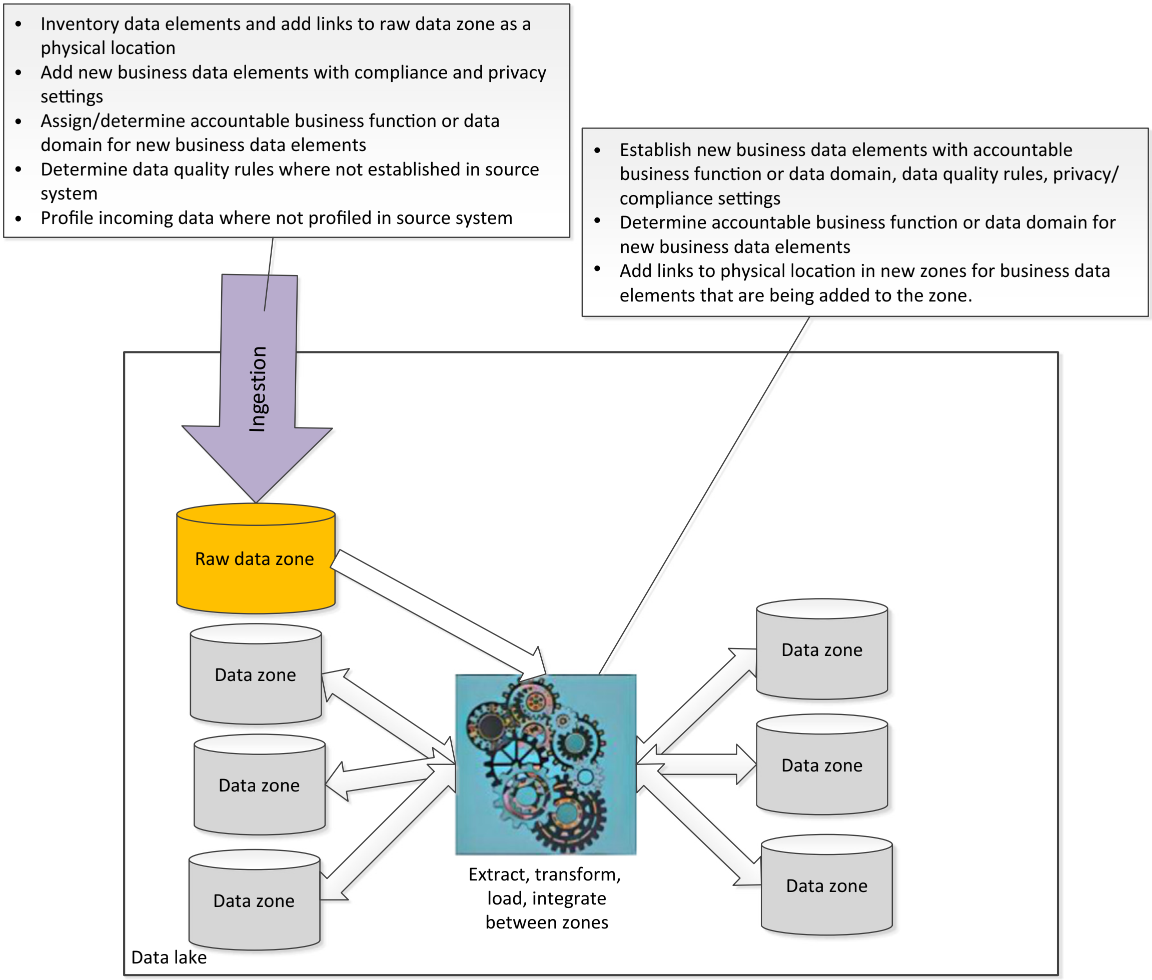
Proposed Roles for Governing in a Data Lake

Due to the complexity of data manipulation in a data lake, there are a lot of roles that get involved with deciding what data to ingest, determining what that data means, how sensitive it is, tracking the lineage, and in gauging the quality. Many of these roles are part of Data Stewardship and Data Governance or work directly with roles in Data Stewardship and Data Governance. Further, as data is moved through various zones in the data lake, and combined and integrated to meet new business needs, Business Data Stewards play increasingly important roles.

Interestingly, most of these roles are *not* new. We have discussed virtually all of them earlier in this book or will discuss them in Chapter 11, Governing and Stewarding Your Data Using Data Domains. Fig. 10.3 shows these roles being involved in common data lake−related tasks such as declaring a new zone, ingesting new data into the lake, and transferring data between zones within the lake.

Data Stewardship in the World of Data Lakes and Fast Data

As with “big data” (discussed previously), there is not much difference between performing Data Stewardship in the new world of data lakes, with just a few exceptions:

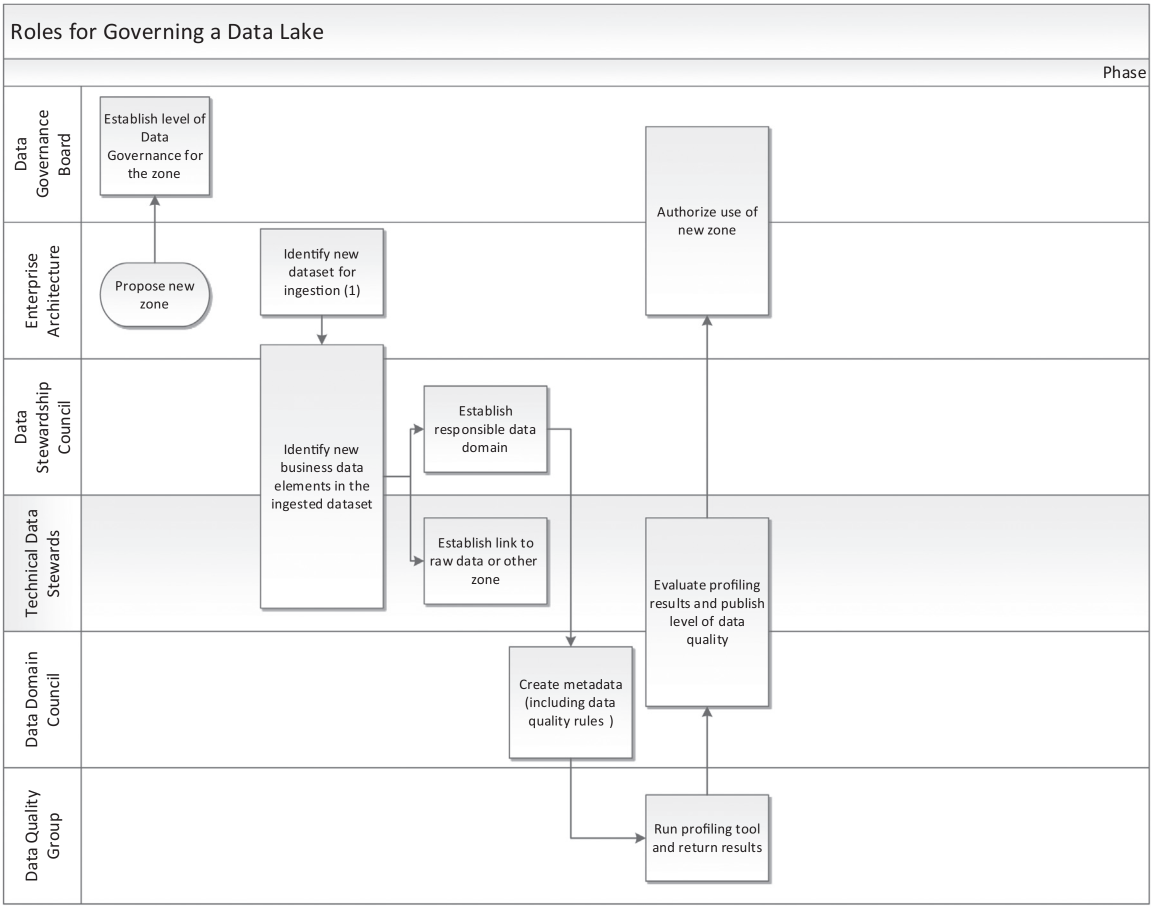


- Inventory data elements and add links to raw data zone as a physical location
- Add new business data elements with compliance and privacy settings
- Assign/determine accountable business function or data domain for new business data elements
- Determine data quality rules where not established in source system
- Profile incoming data where not profiled in source system

- Establish new business data elements with accountable business function or data domain, data quality rules, privacy/compliance settings
- Determine accountable business function or data domain for new business data elements
- Add links to physical location in new zones for business data elements that are being added to the zone.

■ **FIGURE 10.2** Generating metadata as a data lake is evolved and governed.

1. It must be *fast*. With extremely fast changes in what data is ingested, how it is manipulated, and how it is accessed; Data Stewardship cannot afford to be the bottleneck, and the response needs to be quick and efficient—and consistent with the decisions about the level of Data Governance for different zones.
2. It is even *more* important. With the perceived flexibility and ease of exploration of a data lake, the usage is probably going to escalate quickly. But bad decisions can be made if the data is misinterpreted, misunderstood, or of poor quality. Thus governing that data becomes


Roles for Governing a Data Lake
Phase
Data Governance Board
Enterprise Architecture
Data Stewardship Council
Technical Data Stewards
Data Domain Council
Data Quality Group
Establish level of Data Governance for the zone
Propose new zone
Identify new dataset for ingestion (1)
Identify new business data elements in the ingested dataset
Establish responsible data domain
Establish link to raw data or other zone
Create metadata (including data quality rules)
Run profiling tool and return results
Evaluate profiling results and publish level of data quality
Authorize use of new zone

■ **FIGURE 10.3** Roles for some of the governance in a data lake when using data domain-based Data Stewardship.

more critical the more it is used. Further, with many streams of data all converging on the data lake, certain metadata—like lineage—is crucial.

With vast quantities of data and extremely high-powered computing and cheap storage, you can get into a world of "fast data." Fast data handles and uses the data as fast as it arrives, opening new vistas of discovery. Essentially, the data is ingested at millions of events per second and as fast it is ingested, data-driven decisions are made and real-time analyses carried out. But with automated decision-making comes considerable responsibility—many of the scenarios for the use of fast data (such as decisions based on the feed from security cameras) have significant consequences if done incorrectly.

SUMMARY

While there are differences in performing Data Governance and Data Stewardship with big data, fast data, and data lakes, those differences are not large, and do not require a wholesale restructuring or discarding of tried and true procedures. There *are* differences—some decisions need to be made that were not needed before (such as the level of Data Governance for a data lake zone), but most of the differences have to do with a "need for speed" and added complexity due to the nature of a data lake. In addition, with many people needing to contribute knowledge about the data in the data lake, a good case can be made for governing your data by "data domains," as discussed in Chapter 11, Governing and Stewarding Your Data Using Data Domains.