O'REILLY®

Compliments of Zaloni

Revised for 2018

Architecting Data Lakes

Data Management Architectures for Advanced Business Use Cases

Ben Sharma

SECOND EDITION

Architecting Data Lakes

Data Management Architectures for Advanced Business Use Cases

Ben Sharma

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Architecting Data Lakes
by Ben Sharma

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com/safari*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Rachel Roumeliotis
Production Editor: Nicholas Adams
Copyeditor: Octal Publishing, Inc.
Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Rebecca Demarest

March 2016: First Edition
March 2018: Second Edition

Revision History for the Second Edition
2018-02-28: First Release

978-1-492-03297-7

[LSI]

Table of Contents

CHAPTER 1
Overview

Organizations today are bursting at the seams with data, including existing databases, output from applications, and streaming data from ecommerce, social media, apps, and connected devices on the Internet of Things (IoT).

We are all well versed on the data warehouse, which is designed to capture the essence of the business from other enterprise systems—for example, customer relationship management (CRM), inventory, and sales transactions systems—and which allows analysts and business users to gain insight and make important business decisions from that data.

But new technologies, including mobile, social platforms, and IoT, are driving much greater data volumes, higher expectations from users, and a rapid globalization of economies.

Organizations are realizing that traditional technologies can't meet their new business needs.

As a result, many organizations are turning to scale-out architectures such as data lakes, using Apache Hadoop and other big data technologies. However, despite growing investment in data lakes and big data technology—$150.8 billion in 2017, an increase of 12.4% over 2016[1]—just 14% of organizations report ultimately

1 IDC. "Worldwide Semiannual Big Data & Analytics Spending Guide." March 2017.

deploying their big data proof-of-concept (PoC) project into production.[2]

One reason for this discrepancy is that many organizations do not see a return on their initial investment in big data technology and infrastructure. This is usually because those organizations fail to do data lakes right, falling short when it comes to designing the data lake properly and in managing the data within it effectively. Ultimately these organizations create data "swamps" that are really useful for only ad hoc exploratory use cases.

For those organizations that do move beyond a PoC, many are doing so by merging the flexibility of the data lake with some of the governance and control of a traditional data warehouse. This is the key to deriving significant ROI on big data technology investments.

Succeeding with Big Data

The first step to ensure success with your data lake is to design it with future growth in mind. The data lake stack can be complex, and requires decisions around storage, processing, data management, and analytics tools.

The next step is to address management and governance of the data within the data lake, also with the future in mind. How you manage and govern data in a discovery sandbox might not be challenging or critical, but how you manage and govern data in a production data lake environment, with multiple types of users and use cases, is critical. Enterprises need a clear view of lineage and quality for all their data.

It is critical to have a robust set of capabilities to ingest and manage the data, to store and organize it, prepare and analyze it, and secure and govern it. This is essential no matter what underlying platform you choose—whether streaming, batch, object storage, flash, in-memory, or file—you need to provide this consistently through all the evolutions the data lake is going to undergo over the next few years.

The key takeaway? Organizations seeing success with big data are not just dumping data into cheap storage. They are designing and

2 Gartner. "Market Guide for Hadoop Distributions." February 1, 2017.

deploying data lakes for scale, with robust, metadata-driven data management platforms, which give them the transparency and control needed to benefit from a scalable, modern data architecture.

Definition of a Data Lake

There are numerous views out there on what constitutes a data lake, many of which are overly simplistic. At its core, a data lake is a central location in which to store all your data, regardless of its source or format. It is typically built using Hadoop or another scale-out architecture (such as the cloud) that enables you to cost-effectively store significant volumes of data.

The data can be structured or unstructured. You can then use a variety of processing tools—typically new tools from the extended big data ecosystem—to extract value quickly and inform key organizational decisions.

Because all data is welcome, data lakes are a powerful alternative to the challenges presented by data integration in a traditional Data Warehouse, especially as organizations turn to mobile and cloud-based applications and the IoT.

Some of the technical benefits of a data lake include the following:

The kinds of data from which you can derive value are unlimited.
: You can store all types of structured and unstructured data in a data lake, from CRM data to social media posts.

You don't need to have all the answers upfront.
: Simply store raw data—you can refine it as your understanding and insight improves.

You have no limits on how you can query the data.
: You can use a variety of tools to gain insight into what the data means.

You don't create more silos.
: You can access a single, unified view of data across the organization.

The Differences Between Data Warehouses and Data Lakes

The differences between data warehouses and data lakes are significant. A data warehouse is fed data from a broad variety of enterprise applications. Naturally, each application's data has its own schema. The data thus needs to be transformed to be compatible with the data warehouse's own predefined schema.

Designed to collect only data that is controlled for quality and conforming to an enterprise data model, the data warehouse is thus capable of answering a limited number of questions. However, it is eminently suitable for enterprise-wide use.

Data lakes, on the other hand, are fed information in its native form. Little or no processing is performed for adapting the structure to an enterprise schema. The structure of the data collected is therefore not known when it is fed into the data lake, but only found through discovery, when read.

The biggest advantage of data lakes is flexibility. By allowing the data to remain in its native format, a far greater—and timelier—stream of data is available for analysis. Table 1-1 shows the major differences between data warehouses and data lakes.

Table 1-1. Differences between data warehouses and data lakes

Attribute	Data warehouse	Data lake
Schema	Schema-on-write	Schema-on-read
Scale	Scales to moderate to large volumes at moderate cost	Scales to huge volumes at low cost
Access Methods	Accessed through standardized SQL and BI tools	Accessed through SQL-like systems, programs created by developers and also supports big data analytics tools
Workload	Supports batch processing as well as thousands of concurrent users performing interactive analytics	Supports batch and stream processing, plus an improved capability over data warehouses to support big data inquiries from users
Data	Cleansed	Raw and refined
Data Complexity	Complex integrations	Complex processing
Cost/ Efficiency	Efficiently uses CPU/IO but high storage and processing costs	Efficiently uses storage and processing capabilities at very low cost

Attribute	Data warehouse	Data lake
Benefits	• Transform once, use many • Easy to consume data • Fast response times • Mature governance • Provides a single enterprise-wide view of data from multiple sources • Clean, safe, secure data • High concurrency • Operational integration	• Transforms the economics of storing large amounts of data • Easy to consume data • Fast response times • Mature governance • Provides a single enterprise-wide view of data • Scales to execute on tens of thousands of servers • Allows use of any tool • Enables analysis to begin as soon as data arrives • Allows usage of structured and unstructured content form a single source • Supports Agile modeling by allowing users to change models, applications and queries • Analytics and big data analytics
Drawbacks	• Time consuming • Expensive • Difficult to conduct ad hoc and exploratory analytics • Only structured data	• Complexity of big data ecosystem • Lack of visibility if not managed and organized • Big data skills gap

The Business Case for Data Lakes

We've discussed the tactical, architectural benefits of a data lake, now let's discuss the business benefits it provides. Enterprise data warehouses have been most organizations' primary mechanism for performing complex analytics, reporting, and operations. But they are too rigid to work in the era of big data, where large data volumes and broad data variety are the norms. It is challenging to change data warehouse data models, and field-to-field integration mappings are rigid. Data warehouses are also expensive.

Perhaps more important, most data warehouses require that business users rely on IT to do any manipulation or enrichment of data, largely because of the inflexible design, system complexity, and intolerance for human error in data warehouses. This slows down business innovation.

Data lakes can solve these challenges, and more. As a result, almost every industry has a potential data lake use case. For example, almost any organization would benefit from a more complete and nuanced view of its customers and can use data lakes to capture 360-

degree views of those customers. With data lakes, whether used to augment the data warehouse or replace it altogether, organizations can finally unleash big data's potential across industries.

Let's look at a few business benefits that are derived from a data lake.

Freedom from the rigidity of a single data model

Because data can be unstructured as well as structured, you can store everything from blog postings to product reviews. And the data doesn't need to be consistent to be stored in a data lake. For example, you might have the same type of information in very different data formats, depending on who is providing the data. This would be problematic in a data warehouse; in a data lake, however, you can put all sorts of data into a single repository without worrying about schemas that define the integration points between different data sets.

Ability to handle streaming data

Today's data world is a streaming world. Streaming has evolved from rare use cases, such as sensor data from the IoT and stock market data, to very common everyday data, such as social media.

Fitting the task to the tool

A data warehouse works well for certain kinds of analytics. But when you are using Spark, MapReduce, or other new models, preparing data for analysis in a data warehouse can take more time than performing the actual analytics. In a data lake, data can be processed efficiently by these new paradigm tools without excessive prep work. Integrating data involves fewer steps because data lakes don't enforce a rigid metadata schema. *Schema-on-read* allows users to build custom schemas into their queries upon query execution.

Easier accessibility

Data lakes also solve the challenge of data integration and accessibility that plague data warehouses. Using a scale-out infrastructure, you can bring together ever-larger data volumes for analytics—or simply store them for some as-yet-undetermined future use. Unlike a monolithic view of a single enterprise-wide data model, the data lake allows you to put off modeling until you actually use the data, which creates opportunities for better operational insights and data

discovery. This advantage only grows as data volumes, variety, and metadata richness increase.

Scalability

Big data is typically defined as the intersection between volume, variety, and velocity. Data warehouses are notorious for not being able to scale beyond a certain volume due to restrictions of the architecture. Data processing takes so long that organizations are prevented from exploiting all their data to its fullest extent. Petabyte-scale data lakes are both cost-efficient and relatively simple to build and maintain at whatever scale is desired.

Drawbacks of Data Lakes

Despite the myriad technological and business benefits, building a data lake is complicated and different for every organization. It involves integration of many different technologies and requires technical skills that aren't always readily available on the market—let alone on your IT team. Following are three key challenges organizations should be aware of when working to put an enterprise-grade data lake into production.

Visibility

Unlike data warehouses, data lakes don't come with governance built in, and in early use cases for data lakes, governance was an afterthought—or not a thought at all. In fact, organizations frequently loaded data without attempting to manage it in any way. Although situations still exist in which you might want to take this approach—particularly since it is both fast and cheap—in most cases, this type of data dump isn't optimal and ultimately leads to a data swamp of poor visibility into data type, lineage, and quality and really can't be used confidently for data discovery and analytics. For cases in which the data is not standardized, errors are unacceptable, and the accuracy of the data is of high priority, a data dump will greatly impede your efforts to derive value from the data. This is especially the case as your data lake transitions from an add-on feature to a truly central aspect of your data architecture.

Governance

Metadata is not automatically applied when data is ingested into the data lake. Without the technical, operational, and business metadata

that gives you information about the data you have, it is impossible to organize your data lake and apply governance policies. Metadata is what allows you to track data lineage, monitor and understand data quality, enforce data privacy and role-based security, and manage data life cycle policies. This is particularly critical for organizations in tightly regulated industries.

Data lakes must be designed in such a way to use metadata and integrate the lake with existing metadata tools in the overall ecosystem in order to track how data is used and transformed outside of the data lake. If this isn't done correctly, it can prevent a data lake from going into production.

Complexity

Building a big data lake environment is complex and requires integration of many different technologies. Also, determining your strategy and architecture is complicated: organizations must determine how to integrate existing databases, systems, and applications to eliminate data silos; how to automate and operationalize certain processes; how to broaden access to data to increase an organization's agility; and how to implement and enforce enterprise-wide governance policies to ensure data remains private and secure.

In addition, most organizations don't have all of the skills in-house that are needed to successfully implement an enterprise-grade data lake project, which can lead to costly mistakes and delays.

Succeeding with Big Data

The rest of this book focuses on how to build a successful production data lake that accelerates business insight and delivers true business value. At Zaloni, through numerous data lake implementations, we have constructed a data lake reference architecture that ensures production-grade readiness. This book addresses many of the challenges that companies face when building and managing data lakes.

We discuss why an integrated approach to data lake management and governance is essential, and we describe the sort of solution needed to effectively manage an enterprise-grade lake. The book also delves into best practices for consuming the data in a data lake. Finally, we take a look at what's ahead for data lakes.

CHAPTER 2

Designing Your Data Lake

Determining what technologies to employ when building your data lake stack is a complex undertaking. You must consider storage, processing, data management, and so on. Figure 2-1 shows the relationships among these tasks.

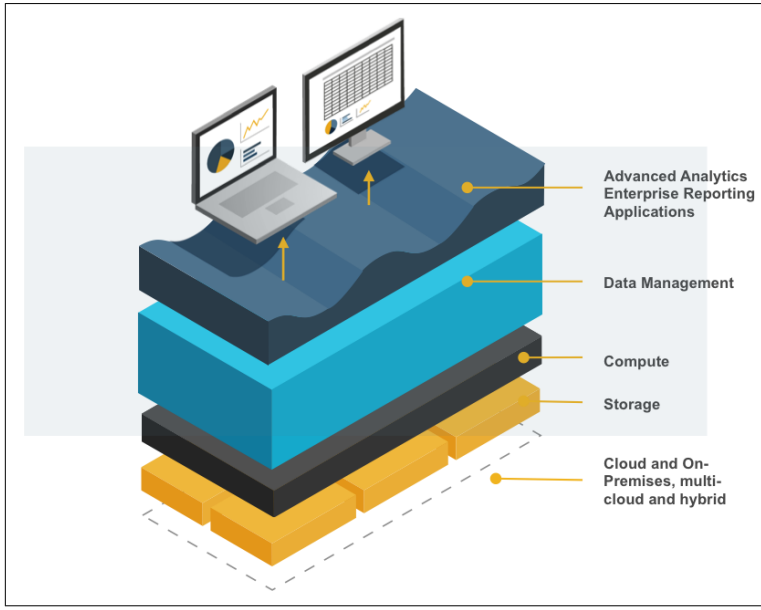

Figure 2-1. The data lake technology stack

Cloud, On-Premises, Multicloud, or Hybrid

In the past, most data lakes resided on-premises. This has undergone a tremendous shift recently, with most companies looking to the cloud to replace or augment their implementations.

Whether to use on-premises or cloud storage and processing is a complicated and important decision point for any organization. The pros and cons to each could fill a book and are highly dependent on the individual implementation. Generally speaking, on-premises storage and processing offers tighter control over data security and data privacy, whereas public cloud systems offer highly scalable and elastic storage and computing resources to meet enterprises' need for large scale processing and data storage without having the overheads of provisioning and maintaining expensive infrastructure.

Also, with the rapidly changing tools and technologies in the ecosystem, we have also seen many examples of cloud-based data lakes used as the incubator for dev/test environments to evaluate all the new tools and technologies at a rapid pace before picking the right one to bring into production, whether in the cloud or on-premises.

If you put a robust data management structure in place, one that provides complete metadata management, you can enable any combination of on-premises storage, cloud storage, and multicloud storage easily.

Data Storage and Retention

A data lake by definition provides much more cost-effective data storage than a data warehouse. After all, with traditional data warehouses' *schema-on-write* model, data storage is highly inefficient—even in the cloud.

Large amounts of data can be wasted due to the data warehouse's *sparse table* problem. To understand this problem, imagine building a spreadsheet that combines two different data sources, one with 200 fields and the other with 400 fields. To combine them, you would need to add 400 new columns into the original 200-field spreadsheet. The rows of the original would possess no data for those 400 new columns, and rows from the second source would hold no data from the original 200 columns. The result? Wasted disk space and extra processing overhead.

A data lake minimizes this kind of waste. Each piece of data is assigned a cell, and because the data does not need to be combined at ingest, no empty rows or columns exist. This makes it possible to store large volumes of data in less space than would be required for even relatively small conventional databases.

In addition to needing less storage, when storage and computing are separate, customers can pay for storage at a lower rate, regardless of computing needs. Cloud service providers like Amazon Web Services (AWS) even offer a range of storage options at different price points, depending on your accessibility requirements.

When considering the storage function of a data lake, you can also create and enforce policy-based data retention. For example, many organizations use Hadoop as an active-archival system so that they can query old data without having to go to tape. However, space becomes an issue over time, even in Hadoop; as a result, there has to be a process in place to determine how long data should be preserved in the raw repository, and how and where to archive it.

A sample technology stack for the storage function of a data lake may consist of the following:

Hadoop Distributed File System (HDFS)
: A Java-based filesystem that provides scalable and reliable data storage. It is designed to span large clusters of commodity servers. For on-premises data lakes, HDFS seems to be the storage of choice because it is highly reliable, fault tolerant, scalable, and can store structured and unstructured data. This allows for faster processing of the big data use-cases. HDFS also allows enterprises to create storage tiers to allow for data life cycle management, using those tiers to save costs while maintaining data retention policies and regulatory requirements.

Object storage
: Object stores (Amazon Simple Storage Service [Amazon S3], Microsoft Azure Blob Storage, Google Cloud Storage) provide scalable, reliable data storage. Cloud-based storage offers a unique advantage. They are designed to decouple storage from computing so that they can autoscale compute power to meet the real-time processing needs.

Apache Hive tables
: An open source data warehouse system for querying and analyzing large datasets stored in Hadoop files.

HBase
: An open source, nonrelational, distributed database that is modeled after Google's BigTable. Developed as part of Apache Software Foundation's Apache Hadoop project, it runs on top of HDFS, providing BigTable-like capabilities for Hadoop.

ElasticSearch
: An open source, RESTful search engine built on top of Apache Lucene and released under an Apache license. It is Java-based and can search and index document files in diverse formats.

Data Lake Processing

Processing transforms data into a standardized format useful to business users and data scientists. It's necessary because during the process of ingesting data into a data lake, the user does not make any decisions about transforming or standardizing the data. Instead, this is delayed until the user reads the data. At that point, the business users have a variety of tools with which to standardize or transform the data.

One of the biggest benefits of this methodology is that different business users can perform different standardizations and transformations depending on their unique needs. Unlike in a traditional data warehouse, users aren't limited to just one set of data standardizations and transformations that must be applied in the conventional schema-on-write approach. At this stage, you can also provision workflows for repeatable data processing.

Appropriate tools can process data for both batch and near-real-time use cases. Batch processing is for traditional extract, transform, and load (ETL) workloads—for example, you might want to process billing information to generate a daily operational report. Streaming is for scenarios in which the report needs to be delivered in real time or near real time and cannot wait for a daily update. For example, a large courier company might need streaming data to identify the current locations of all its trucks at a given moment.

Different tools are needed, based on whether your use case involves batch or streaming. For batch use cases, organizations generally use

Pig, Hive, Spark, and MapReduce. For streaming use cases, they would likely use different tools such as Spark-Streaming, Kafka, Flume, and Storm.

A sample technology stack for processing might include the following:

MapReduce
: MapReduce has been central to data lakes because it allows for distributed processing of large datasets across processing clusters for the enterprise. It is a programming model and an associated implementation for processing and generating large datasets with a parallel, distributed algorithm on a cluster. You can also deploy it on-premises or in a cloud-based data lake to allow a hybrid data lake using a single distribution (e.g., Cloudera, Hortonworks, or MapR).

Apache Hive
: This is a mechanism to project structure onto large datasets and to query the data using a SQL-like language called HiveQL.

Apache Spark
: Apache Spark is an open source engine developed specifically for handling large-scale data processing and analytics. It provides a faster engine for large-scale data processing using in-memory computing. It can run on Hadoop, Mesos, in cloud, or in a standalone environment to create a unified compute layer across the enterprise.

Apache Drill
: An open source software framework that supports data-intensive distributed applications for interactive analysis of large-scale datasets.

Apache Nifi
: This is a framework to automate the flow of data between systems. NiFi's Flow-Based Programming (FBP) platform allows data processing pipelines to address end-to-end data flow in big-data environments.

Apache Beam
: Apache Beam provides an abstraction on top of the processing cluster. It is an open source framework that allows you to use a single programming model for both batch and streaming use

cases, and execute pipelines on multiple execution environments like Spark, Flink, and others. By utilizing Beam, enterprises can develop their data processing pipelines using Beam SDK and then choose a Beam Runner to run the pipelines on a specific large-scale data processing system. The runner can be a number of things: a Direct Runner, Apex, Flink, Spark, Dataflow, or Gearpump (incubating). This design allows for the processing pipeline to be portable across different runners, thereby providing flexibility to the enterprises to take advantage of the best platform to meet their data processing requirements in a future-proof way.

Data Lake Management and Governance

At this layer, enterprises need tools to ingest and manage their data across various storage and processing layers while maintaining clear track of data throughout its life cycle. This not only provides an efficient and fast way to derive insights, but also allows enterprises to meet their regulatory requirements around data privacy, security, and governance.

Data lakes created with an integrated data management framework can eliminate the cumbersome data preparation process of ETL that traditional data warehouse requires. Data is smoothly ingested into the data lake, where it is managed using metadata tags that help locate and connect the information when business users need it. This approach frees analysts for the important task of finding value in the data without involving IT in every step of the process, thus conserving IT resources. Today, all IT departments are being mandated to do more with less. In such environments, well-managed data lakes help organizations more effectively utilize all of their data to derive business insight and make good decisions.

Data governance is critically important, and although some of the tools in the big data stack offer partial data governance capabilities, organizations need more advanced capabilities to ensure that business users and data scientists can track data lineage and data access, and take advantage of common metadata to fully make use of enterprise data resources.

Key to a solid data management and governance strategy is having the right metadata management structure in place. With accurate and descriptive metadata, you can set policies and standards for

managing and using data. For example, you can create policies that enforce users' ability to acquire data from certain places, which these users then own and are therefore responsible; which users can access the data; how the data can be used and how it's protected—including how it is stored, archived, and backed up.

Your governance strategy must also specify how data will be audited to ensure that you are complying with government regulations that apply to your industry (sometimes on an international scale, such as the European Union's General Data Protection Regulation [GDPR]). This can be tricky to control while diverse datasets are combined and transformed. All of this is possible if you deploy a robust data management platform that provides the technical, operational, and business metadata required.

Advanced Analytics and Enterprise Reporting

This stage is where the data is consumed from the data lake. There are various modes of accessing the data: queries, tool-based extractions, or extractions that need to happen through an API. Some applications need to source the data for performing analyses or other transformations downstream.

Visualization is an important part of this stage, where the data is transformed into charts and graphics for easier comprehension and consumption. Tableau and Qlik are two popular tools offering effective visualization. Business users can also use dashboards, either custom-built to fit their needs, or off-the-shelf such as Microsoft SQL Server Reporting Services, Oracle Business Intelligence Enterprise Edition, or IBM Cognos.

Application access to the data is provided through APIs, message-queues, and database access.

Here's an example of what your technology stack might look like at this stage:

Qlik
: Allows you to create visualizations, dashboards, and apps that answer important business questions.

Tableau
: Business intelligence software that allows users to connect to data, and create interactive and shareable dashboards for visualization.

Spotfire
: Data visualization and analytics software that helps users quickly uncover insights for better decision making.

RESTful APIs
: An API that uses HTTP requests to GET, PUT, POST, and DELETE data.

Apache Kafka
: A fast, scalable, durable, and fault-tolerant publish–subscribe messaging system, Kafka brokers massive message streams for low-latency analysis in Enterprise Apache Hadoop.

Java Database Connectivity (JDBC)
: An API for the programming language Java, which defines how a client can access a database. It is part of the Java Standard Edition platform, from Oracle Corporation.

The Zaloni Data Lake Reference Architecture

A reference architecture is a framework that organizations can refer to in order to 1) understand industry best practices, 2) track a process and the steps it takes, 3) derive a template for solutioning, and 4) understand the components and technologies involved.

Our reference architecture has less to do with how the data lake fits into the larger scheme of a big data environment, and more to do with how the data lake is managed. Describing how the data will move and be processed through the data lake is crucial to understanding the system as well as making it more user friendly. Furthermore, it provides a description of the capabilities a well-managed and governed data lake can and should have, which can be taken and applied to a variety of use cases and scenarios.

We recommend organizing your data lake into four zones, plus a sandbox, as illustrated in Figure 2-2. Throughout the zones, data is tracked, validated, cataloged, assigned metadata, refined, and more. These capabilities and the zones in which they occur help users and moderators understand what stage the data is in and what measures

have been applied to them thus far. Users can access the data in any of these zones, provided they have appropriate role-based access.

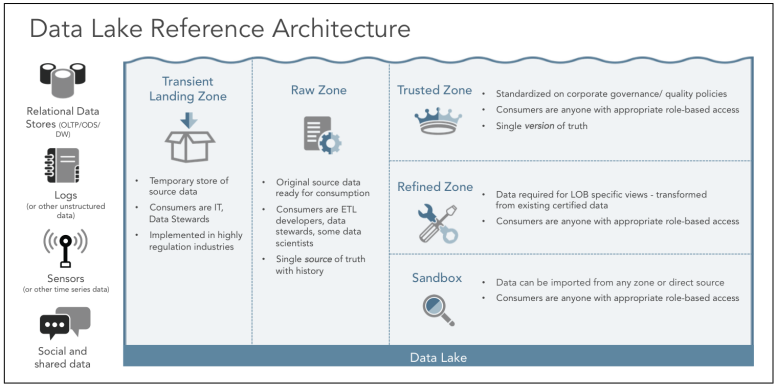

Figure 2-2. The Zaloni data lake reference architecture outlines best practices for storing, managing, and governing data in a data lake

Data can come into the data lake from anywhere, including online transaction processing (OLTP) or operational data store (ODS) systems, a data warehouse, logs or other machine data, or from cloud services. These source systems include many different formats, such as file data, database data, ETL, streaming data, and even data coming in through APIs.

Zone 1: The Transient Landing Zone

We recommend loading data into a transient loading zone, where basic data quality checks are performed using MapReduce or Spark processing capabilities. Many industries require high levels of compliance, with data having to pass a series of security measures before it can be stored. This is especially common in the finance and healthcare industries, for which customer information must be encrypted so that it cannot be compromised. In some cases, data must be masked prior to storage.

The transient zone is temporary; it is a landing zone for data where security measures can be applied before it is stored or accessed. With GDPR being enacted within the next year in the EU, this zone might become even more important because there will be higher levels of regulation and compliance, applicable to more industries.

Zone 2: The Raw Zone

After the quality checks and security transformations have been performed in the Transient Zone, the data is then loaded into in the Raw Data zone for storage. However, in some situations, a Transient Zone is not needed, and the Raw Zone is the beginning of the data lake journey.

Within this zone, you can mask or tokenize data as needed, add it to catalogs, and enhance it with metadata. In the Raw Zone, data is stored permanently and in its original form, so it is known as "the single source of truth." Data scientists and business analysts alike can dip into this zone for sets of data to discover.

Zone 3: The Trusted Zone

The Trusted Zone imports data from the Raw Zone. This is where data is altered so that it is in compliance with all government and industry policies as well as checked for quality. Organizations perform standard data cleansing and data validation methods here.

The Trusted Zone is based on raw data in the Raw Zone, which is the "single source of truth." It is altered in the Trusted Zone to fit business needs and be in accordance with set policies. Often the data within this zone is known as a "single version of truth."

This trusted repository can contain both master data and reference data. Master data is a compilation of the basic datasets that have been cleansed and validated. For example, a healthcare organization might have master data that contains basic member information (names, addresses) and members' additional attributes (dates of birth, social security numbers). An organization needs to ensure that data kept in the trusted zone is up to date using change data capture (CDC) mechanisms.

Reference data, on the other hand, is considered the single version of truth for more complex, blended datasets. For example, that healthcare organization might have a reference dataset that merges information from multiple source tables in the master data store, such as the member basic information and member additional attributes, to create a single version of truth for member data. Anyone in the organization who needs member data can access this reference data and know they can depend on it.

Zone 4: The Refined Zone

Within the Refined Zone, data goes through its last few steps before being used to derive insights. Data here is integrated into a common format for ease of use, and goes through possible detokenization, further quality checks, and life cycle management. This ensures that the data is in a format from which you can easily use it to create models. Consumers of this zone are those with appropriate role-based access.

Data is often transformed to reflect the needs of specific lines of business in this zone. For example, marketing streams might need to see the ROI of certain engagements to gauge their success, whereas finance departments might need information displayed in the form of balance sheets.

The Sandbox

The Sandbox is integral to a data lake because it allows data scientists and managers to create ad hoc exploratory use cases without the need to involve the IT department or dedicate funds to creating suitable environments within which to test the data.

Data can be imported into the Sandbox from any of the zones, as well as directly from the source. This allows companies to explore how certain variables could affect business outcomes and therefore derive further insights to help make business management decisions. You can send some of these insights directly back to the raw zone, allowing derived data to act as sourced data and thus giving data scientists and analysts more with which to work.

CHAPTER 3
Curating the Data Lake

Although it is exciting to have a cost-effective scale-out platform, without controls in place, no one will trust it for business-critical applications. It might work for ad hoc use cases, but you still need a management and governance layer that organizations are accustomed to having in traditional data warehouse environments if you want to scale and use the value of the lake.

For example, consider a bank aggregating risk data across different lines of business into a common risk reporting platform for the Basel Committee on Banking Supervision (BCBS) 239. The data must be of very high quality and have good lineage to ensure that the reports are correct, because the bank depends on those reports to make key decisions about how much capital to carry. Without this lineage, there are no guarantees that the data is accurate.

A data lake makes perfect sense for this kind of data, because it can scale out as you bring together large volumes of different risk datasets across different lines of business. But data lakes need a management platform in order to support metadata as well as quality and governance controls. To succeed at applying data lakes to these kinds of business use cases, you need controls in place.

This includes the right tools and the right process. Process can be as simple as assigning stewards to new datasets, or forming a data lake enterprise data council, to establish data definitions and standards.

Questions to ask when considering goals for data governance include the following:

Quality and consistency
: What is needed to ensure that the data is of sufficient quality and consistency to be useful to business users and data scientists in making important discoveries and decisions?

Policies and standards
: What are the policies and standards for ingesting, transforming, and using data, and are they observed uniformly throughout the organization?

Security, privacy, and compliance
: Is access to sensitive data limited to those with the proper authorization?

Data life cycle management
: How will we manage the life cycle of the data? At what point will we move it from expensive Tier-1 storage to less-expensive storage mechanisms?

Integrating Data Management

We believe that effective data management and governance is best delivered through an integrated platform, such as the Zaloni Data Platform (ZDP). The alternative is to perform the best practices from the previous section in siloes, thereby wasting a large amount of time stitching together different point products. You would end up spending a great deal of resources on the plumbing layer of the data lake—the platform—when you could be spending resources on something of real value to the business such as analyses and insights your business users gain from the data.

Having an integrated platform improves your time-to-market for insights and analytics tremendously, because all of these aspects fit together. As you ingest data, the metadata is captured. As you transform the data into a refined form, lineage is automatically captured. Rules ensure that all incoming data is inspected for quality—so whatever data you make available for consumption goes through these data quality checks.

An effective way to discuss the many components of data management and governance is to look at it in order of a typical data pipeline. At Zaloni, we look at the stages along the pipeline from data source to data consumer as Ingest, Organize, Enrich, Engage. In the sections that follow, we discuss these areas in detail and also look at

automation, workflow orchestration and other ways in which to operationalize your data lake.

Data Ingestion

Organizations have a number of options when transferring data to a data lake. Managed ingestion gives you control over how data is ingested, where it comes from, when it arrives, and where it is stored in the data lake.

With managed ingestion, data (both structured and unstructured) and metadata are linked. There are two approaches to link data and metadata. The first approach is when metadata is known; for example, a meta file that describes the column name and data type that comes with the data file.

The second approach is when metadata is not known ahead of time and the data stewards identify and tag individual cells after data is ingested. With this approach, organizations can handle both structured data that comes from CSV or RDBMS and unstructured data such as photographs, Twitter feeds, or blog posts and derive business value from the data being collected. This overcomes the traditional schema-on-write limitation of traditional data warehouses.

With managed ingestion, you enter all data into a giant table organized with metadata tags. Each piece of data—whether a customer's name, a photograph, or a Facebook post—is placed in an individual cell. It doesn't matter where in the data lake that individual cell is located, where the data came from, or its format. You can connect all of the data easily through the tags. You can add or change tags as your analytic requirements evolve—one of the key distinctions between a data warehouse and a data lake.

Using a managed ingestion process with a data lake opens up tremendous possibilities. You can quickly and easily ingest unstructured data and make it available for analysis without needing to transform it in any way.

You can also protect sensitive information. As data is ingested into the data lake and moves from the Transient Zone to the Raw Zone, each cell is tagged according to how "visible" it is to different users in the organization. In other words, you can specify who has access to the data in each cell, and under what circumstances, right from the beginning of ingestion.

For example, a retail operation might make cells containing customers' names and contact data available to employees in sales and customer service, but it might make the cells containing more sensitive personally identifiable information (PII) or financial data available only to personnel in the finance department. That way, when users run queries on the data lake, their access rights restrict the visibility of the data.

A key benefit of managed ingestion is that it gives IT the tools to troubleshoot and diagnose ingestion issues before they become problems. For example, with Zaloni's Data Lake Management Platform, all steps of the data ingestion pipeline are defined in advance, tracked, and logged; the process is repeatable and scalable. The platform also simplifies the onboarding of new data sets and can ingest from files, databases, streaming data, REST APIs, and cloud storage services like Amazon Simple Storage Service (Amazon S3).

An integrated data lake management platform will perform managed ingestion, which involves getting the data from the source systems into the data lake, guaranteeing a repeatable process, and ensuring that if anything fails in the ingest cycle, there will be operational functions that take care of it.

For example, a platform implementing managed ingestion can raise notifications and captures logs, so that you can debug why an ingestion failed, fix it, and restart the process. This is all tied with post-processing after the data is stored in the data lake.

Additionally, as we see more and more workloads moving to streaming, whatever data management functions you applied to batch ingestion—when data was coming in periodically—now needs to be applied to data that is streaming in continuously. Integrated data lake management platforms should be able to detect whether certain streams are not being ingested based on the service-level agreements (SLAs) you set.

A data management platform should ensure that the capabilities available in the batch ingestion layer are also available in the streaming ingestion layer. Metadata still needs to be captured and data quality checks need to be performed for streaming data. And you still need to validate that the record format is correct, and that the record values are correct by doing range checks or reference integrity checks.

A data management solution that is purpose-built to provide these capabilities provides the foundation for a well-defined data pipeline. Of course, you need the right business processes, too, such as assigning stewards for new datasets that are ingested.

The basic requirements when ingesting data into the data lake include the following:

1. Define the incoming data from a business perspective.
2. Document the context, lineage, and frequency of the incoming data.
3. Classify the security level (public, internal, sensitive, restricted) of the incoming data.
4. Document the creation, usage, privacy, regulatory, and encryption business rules that apply to the incoming data.
5. Identify the data owner (sponsor) of the ingested data.
6. Identify the data steward(s) charged with monitoring the health of the specific data sets.
7. Continuously measure data quality as it resides in the data lake.

Data Governance

An important part of the data lake architecture is to first put data in a transient area before moving it to the raw data repository. From this staging area, all possible data sources, external or internal, are either moved into the data lake or discarded. As with the visibility of the data, a managed ingestion process enforces governance rules that apply to all data that is allowed to enter the data lake.

Governance rules can include any or all of the following:

Encryption
: If data needs to be protected by encryption—if its visibility is a concern—it must be encrypted before it enters the data lake.

Provenance and lineage
: It is particularly important for the analytics applications that business analysts and data scientists will use down the road that the data provenance and lineage is recorded. You might even want to create rules to prevent data from entering the data lake if its provenance is unknown.

Metadata capture
: A managed ingestion process allows you to set governance rules that capture the metadata on all data before it enters the data lake's raw repository.

Data cleansing
: You can also set data cleansing standards that are applied as the data is ingested in order to ensure that only clean data makes it into the data lake.

You can approach these governances from different angles and methodologies. A top-down method takes best practices from organizations' data warehouse experiences and attempts to impose governance and management from the moment the data is ingested into the data lake. Other solutions take a bottom-up approach that allows users to explore, discover, and analyze the data much more fluidly and flexibly.

We recommend weaving together a combined approach that is specifically suited for your organization that exploits the benefits from the top-down and bottom-up processes. For example, some top-down process is essential if the data from the data lake is going to be a central part of the enterprise's overall data architecture. At the same time, much of the data lake can be managed from the bottom up, including managed data ingestion, data inventory, data enrichment, data quality, metadata management, data lineage, workflow, and self-service access.

With a top-down approach, data governance policies are defined by a centralized body within the organization, such as a chief data officer's office, and are enforced by all of the different functions as they build out the data lake. Policies cover data quality, data security, source systems that can provide data, the frequency of the updates, the definitions of the metadata, identifying the critical data elements, and centralized processes driven by a centralized data authority.

In a bottom-up approach, consumers of the data lake are probably data scientists or data analysts. Collective input from these consumers decides which datasets are valuable and useful and have good quality data. The data lake then exposes those datasets to other consumers so that they can see the ways that their peers have been successful with the data lake.

With a combined approach, you avoid hindering agility and innovation (which can happen with the top-down approach), and at the same time, you avoid the chaos risked by a completely bottom-up approach.

Data Catalog

With the distributed Hadoop Distributed File System (HDFS) filesystem, information is first broken up into blocks, and then written in a distributed manner in the cluster. However, sometimes you need to see metadata such as what datasets exist in the data lake, the properties of those datasets, the ingestion history of the dataset, the data quality, and the key performance indicators (KPIs) of the data as it was ingested. You should also see the data profile and all the metadata attributes, including those that are business, technical, and operational. All of these things need to be abstracted to a level at which the user can understand them and use that data effectively. This is where the data lake catalog comes in.

Your management platform should make it easy to create a data catalog and provide it to business users, so they can easily search it, whether searching for source system, schema attributes, subject area, or time range. This is essential if your business users are to get the most out of the data lake, and use it in a swift and agile way.

With a data catalog, users can find datasets that are curated so that they don't spend time cleaning up and preparing the data. This has already been done for them, particularly in cases of data that has made it to the trusted area. Users are thus able to select the datasets that they want for model building without involving IT, which shortens the analytics timeline.

Capturing Metadata

Metadata is extraordinarily important to managing your data lake. Metadata is critical for making sure data is used to its fullest. Whether manually collected or automatically created during data ingestion, metadata allows your users to locate the data that they want to analyze. It also provides clues for future users to understand the contents of a dataset and how it could be reused.

An integrated data lake management platform makes metadata creation and maintenance an integral part of the data lake processes.

This is essential, as without effective metadata, data dumped into a data lake might never be seen again.

A lot of requirements for metadata can be defined by your organization's central data authority, by your chief data officer, or by data stewards in your lines of business, who might want to specify the various attributes and entities of data that they are bringing into the data lake.

As data lakes grow deeper and more important to the conduct of daily business, metadata is a vital tool in ensuring that you can find the data you pour into these lakes and harness it for years to come. There are three distinct but equally important types of metadata to collect: technical, operational, and business data, as shown in Table 3-1.

Table 3-1. Three equally important types of metadata

Type of metadata	Description	Example
Technical	Captures the form and structure of each dataset	Type of data (text, JSON, Avro); structure of the data (the fields and their types)
Operational	Captures lineage, quality, profile, and provenance of the data	Source and target locations of data, size, number of records, and lineage
Business	Captures what it all means to the user	Business names, destinations, tags, quality, and masking rules

Technical metadata captures the form and structure of each dataset. For example, it captures the type of data file (text, JSON, Avro) and the structure of the data (the fields and their types), and other technical attributes. This is either automatically associated with a file upon ingestion or discovered manually after ingestion.

Operational metadata captures the lineage, quality, profile, and provenance of the data at both the file and the record levels; the number of records; and the lineage. Someone must manually enter and tag entities with operational metadata.

Business metadata captures what the user needs to know about the data, such as the business names, the descriptions of the data, the tags, the quality, and the masking rules for privacy. All of this can be automatically captured by an integrated data management platform upon ingestion.

Leading integrated data management solutions will possess file- and record-level watermarking features that enable you to see the data lineage, where data moves, and how it is used. These features safeguard data and reduce risk, because the data manager will always know where data has come from, where it is, and how it is being used.

Data Privacy

When you store data, depending on the use case, you might need to consider some security encryption requirements. You might need to mask or tokenize data and protect it with proper access controls.

A core attribute of the data lake architecture is to share centrally stored data among multiple groups. Although this is very efficient, you need to make sure that all users have appropriate permission levels to view the information. For example, in a healthcare organization, certain information is deemed private by law, such as protected health information (PHI), and violators—organizations that don't protect this PHI—are severely penalized.

The data preparation stage is often where sensitive data, such as financial and health information, is protected. An integrated management platform can perform masking (where data from a field is completely removed) and tokenization (changing parts of the data to something innocuous). This type of platform also enforces a policy-based mechanism, like access control lists, to make sure the data is protected appropriately.

Storage Considerations via Data Life Cycle Management

It's important to consider the best format for storing the data. You might need to store it in the raw format in which it came, but you might also want to store it in a format that is more consumable for business users, so that queries will run faster. For example, queries that run on columnar data sets will return much faster results than those in a typical row-structured dataset. You might also want to compress the data because it might be entering the system in large volumes and you want to save on storage.

Also, when storing data, the platform should ideally enable you to automate data life cycle management functions. For example, you might store the data in different zones in the data lake, depending on different SLAs. For example, as raw data comes in, you might want to store that is used very frequently in a "hot zone," for, perhaps, 30 days. Then, you might want to move it to a "warm zone" for 90 days, and from there to a "cold zone" for seven years, from which the queries are much more infrequent.

Data Preparation

To meet the business goals for which the data lake was created, business users must find it easy to access and use the data that resides in the data lake without depending on IT assistance.

However, just adding raw data to the data lake does not make that data ready for use by data and analytics applications: data preparation is required. Inevitably, data will come into the data lake with a certain amount of errors, corrupted formats, or duplicates. A data management platform makes it easier to adequately prepare and clean the data by using built-in functionality that delivers data security, quality, and visibility. Workflow orchestration automatically applies rules for data preparation to new data as it flows into the lake.

Data preparation capabilities of an integrated data lake management platform should include the following:

- Data tagging so that searching and sorting becomes easier
- Converting data formats to make executing queries against the data faster
- Executing complex workflows to integrate updated or changed data

Whenever you do any of these data preparation functions, you need metadata that shows the lineage from a transformation perspective. What queries were run? When did they run? What files were generated? You need to create a lineage graph of all the transformations that happen to the data as it flows through the pipeline.

Additionally, when going from raw to refined, you might want to watermark the data by assigning a unique ID for each record of the data so that you can trace a record back to its original file. You can

watermark at either the record or file level. Similarly, you might need to do format conversions as part of your data preparation; for example, if you prefer to store the data in a columnar format.

Other issues can arise. You might have changes in data coming from source systems. How do you reconcile that changed data with the original datasets you brought in? You should be able to maintain a time series of what happens over a period of time.

A data management platform like ZDP can do all of this and ensure that all necessary data preparation is completed before the data is published into the data lake for consumption.

Benefits of an Integrated Approach

Taking an integrated data management approach to a data lake ensures that each business unit does not build a separate cluster—a common practice with data warehouses. Instead, you build a data lake with a shared enterprise cluster. An integrated management platform provides the governance and the multitenant capabilities to do this, and to implement best practices for governance without a negative impact on the speed or agility of the data lake. This type of platform enables you to do the following:

Understand provenance
: Track the source and lineage of any data loaded into the data lake. This gives you traceability of the data, indicates where it came from, when it came in, how many records it has, and whether the dataset was created from other datasets. These details allow you to establish accountability, and you can use this information to do impact analysis on the data.

Understand context
: Record data attributes, such as the purpose for which the data was collected, the sampling strategies employed in its collection, and any data dictionaries or field names associated with it. These pieces of information make your organization much more productive as you progress along the analytics pipeline to derive insights from the data.

Track updates
: Log each time new data is loaded from the same source and record any changes to the original data introduced during an

update. You need to do this for cases in which data formats keep changing.

For example, suppose that you are working with a retail chain, with thousands of point-of-sale (PoS) terminals sending data from 8,000-plus stores in the United States. These PoS terminals are gradually upgraded to newer versions, but not everything can be upgraded on a given day—and now you have multiple formats of data coming in. How do you keep track as the data comes in? How do you know what version it maps to? How do you associate it with the appropriate metadata and structures so that it can be efficiently used for building the analytics? All of these questions can be answered with a robust integrated data lake management platform.

Track modifications
: Record when data is actively changed, and know by whom and how it was done. If there are format changes, you are able to track them as you go from version 1 to version 2, so you know which version of the data you are processing and the structure or schemes associated with that version.

Perform transformations
: Convert data from one format to another to de-duplicate, correct spelling, expand abbreviations, or add labels. Driven by metadata, these transformations are greatly streamlined. And because they are based on metadata, the transformations can accommodate changes in a much more dynamic manner. For example, you have a record format with 10 fields and perform a transformation based on metadata of these fields. If you decide to add an additional field, you can adjust that transformation without having to start from scratch. In other words, the transformation is driven by and integrated with the metadata.

Track transformations
: Performing transformations is a valuable ability, but an additional, essential requirement involves keeping track of the transformations you have accomplished. With a leading integrated data management platform, you can record the ways in which datasets are transformed. Suppose that you perform a transformation from a source to a target format: you can track the lineage so that you know, for example, that this file and these

records were transformed to a new file in this location and in this format, which now has this many records.

Manage metadata
: Manage all of the metadata associated with all of the previously listed activities, making it easy to track, search, view, and act upon all of your data. Because you are using an integrated approach, much of technical metadata can be discovered from the data coming in, and the operational data can be automatically captured without any manual steps. This capability provides you with a much more streamlined approach for collecting metadata.

CHAPTER 4

Deriving Value from the Data Lake

Self-service consumption is essential for a successful data lake. Different types of users consume the data, and they are looking for different things—but each wants to access the data in a self-service manner, without the help of IT.

The Executive

An executive is usually a person in senior management looking for high-level analyses that can help them make important business decisions. For example, an executive could be looking for predictive analytics of product sales based on history and analytical models built by data scientists. In an integrated data lake management platform, data would be ingested from various sources—some streaming, some batch—and then processed in batches to come up with insights, with the final data able to be visualized using Tableau or Excel. Another common example is an executive who needs a 360-degree view of a customer, including metrics from every level of the organization—pre-sales, sales, and customer support—in a single report.

The Data Scientist

Data scientists are typically looking at the datasets and trying to build models on top of them, performing exploratory ad hoc analyses to prove or come up with a thesis about what they see. Data scientists who want to build and test their models will find a data lake

useful because it gives them access to all of the data, not just a sample. Additionally, they can build scripts in Python and run them on a cluster to get a response within hours rather than days.

The Business Analyst

Business analysts usually try to correlate some of the datasets and create an aggregated view to slice and dice using a business intelligence or visualization tool. With a traditional data warehouse, business analysts had to come up with reporting requirements and wait for IT to build a report or export the data on their behalf. Now, business analysts can ask "what if" questions from data lakes on their own. For example, an analyst might ask how much effect weather patterns had on sales based on historical data and information from public datasets combined with in-house datasets in the data lake. Without involving IT, the analyst could consult the catalog to see what datasets have been cleaned and standardized and run queries against that data.

The Downstream System

A fourth type of consumer is a downstream system, such as an application or a platform, which receives the raw or refined data. Leading companies are building new applications and products on top of their data lake, so they are also consumers of the data. They might also use RESTful APIs or some other API mechanisms on an ongoing manner. For example, if the downstream application is a database, the data lake can ingest and transform the data and then send the final aggregated data to the downstream system for storage.

Self-Service

The purpose of a data lake is to provide value to the business by serving users. From a user perspective, here are the most important questions to ask about the data:

- What is in the data lake (the catalog)?
- What is the quality of the data?
- What is the profile of the data?
- What is the metadata of the data?

- How can users do enrichments, clean-ups, enhancements, and aggregations without going to IT (how to use the data lake in a self-service way)?
- How can users annotate and tag the data?

Answering these questions requires that proper architecture, governance, and security rules are put in place and adhered to so that the appropriate people gain access to the relevant data in a timely manner. There also needs to be strict governance in the onboarding of datasets, naming conventions must be established and enforced, and security policies need to be in place to ensure role-based access control.

For our purposes, self-service means that nontechnical business users can access and analyze data without involving IT. In a self-service model, users should be able to see the metadata and profiles and understand what the attributes of each dataset mean. The metadata must provide enough information for users to create new data formats out of existing data formats by using enrichments and analytics.

Also, in a self-service model, the catalog will be the foundation for users to register all of the different datasets in the data lake. This means that users can go to the data lake and search to find the datasets they need. They should also be able to search on any kind of attribute; for example, on a time window such as January 1st to February 1st, or based on a subject area, such as marketing versus finance. Users should also be able to find datasets based on attributes; for example, they could enter, "Show me all of the datasets that have a field called discount or percentage."

It is in the self-service capability that best practices for the various types of metadata come into play. Business users are interested in the business metadata, such as the source systems, the frequency with which the data comes in, and the descriptions of the datasets or attributes. Users are also interested in knowing the technical metadata: the structure, format, and schema of the data.

When it comes to operational data, users want to see information about lineage, including when the data was ingested into the data lake, and whether it was raw at the time of ingestion. If the data was not raw when ingested, users should be able to see how was it created and what other datasets were used to create it. Also important

to operational data is the quality of the data. Users should be able to define certain rules about data quality, and use them to perform checks on the datasets.

Users might also want to see the ingestion history. If a user is looking at streaming data, for example, they might search for days where no data came in, as a way of ensuring that those days are not included in the representative datasets for campaign analytics. Overall, access to lineage information, the ability to perform quality checks, and ingestion history give business users a good sense of the data, making it possible for them to quickly begin analytics.

Controlling Access

Many IT organizations are simply overwhelmed by the sheer volume of datasets—small, medium, and large—that are related but not integrated when they are stored in data lakes. However, when done right, data lakes allow organizations to gain insights and discover relationships between datasets.

When providing various users—whether C-level executives, business analysts, or data scientists—with the tools they need, security is critical. Setting and enforcing the security policies consistently is essential for successful use of a data lake. In-memory technologies should support different access patterns for each user group, depending on their needs. For example, a report generated for a C-level executive might be very sensitive and should not be available to others who don't have the same access privileges. Data scientists might need more flexibility, with lesser amounts of governance; for this group, you might create a sandbox for exploratory work. By the same token, users in a company's marketing department should not have access to the same data as users in the finance department. With security policies in place, users have access only to the datasets assigned to their privilege levels.

You can also use security features to enable users to interact with the data and contribute to data preparation and enrichment. For example, as users find data in the data lake through the catalog, they can be allowed to clean up the data and enrich the fields in a dataset in a self-service manner.

Access controls can also enable a collaborative approach for accessing and consuming the data. For example, if one user finds a dataset

that is important to a project, and there are three other team members on that same project, the user can create a shared workspace with that data so that the team can collaborate on enrichments.

Crowdsourcing

A bottom-up approach to data governance enables you to rank the usefulness of datasets by crowdsourcing. By asking users to rate which datasets are the most valuable, the word can spread to other users so that they can make productive use of that data.

To do this, you need a rating and ranking mechanism as part of your integrated data lake management platform. The obvious place for this bottom-up, watermark-based governance model would be the catalog. Thus, the catalog must have rating functions.

But it's not enough to show what others think of a dataset. An integrated data lake management and governance solution should show users the rankings of the datasets from all users, but it should also offer a personalized data rating, so that each individual can see what they have personally found useful whenever they go to the catalog.

Users also need tools to create new data models out of existing datasets. For example, users should be able to take a customer data set and a transaction dataset and create a "most valuable customer" dataset by grouping customers by transactions and determining when customers are generating the most revenue. Being able to do these types of enrichments and transformations is important from an end-to-end perspective.

Data Lakes in Different Industries

The data lake provides value in many different areas. Following are some examples industries that benefit from using a data lake to store, transform, and access information.

Health and Life Sciences

Data lakes allow health and life sciences organizations and companies to store and access widely disparate records of both structured and unstructured data in their native formats for later analysis. This avoids the need to force a single categorization of each data type, as would be the case in a traditional data warehouse. Not incidentally,

preserving the native format also helps maintain data provenance and fidelity of the data, enabling different analyses to be performed using different contexts. With data lakes, sophisticated data analysis projects are now possible because the data lakes enable distributed big data processing using broadly accepted, open software standards and massively parallel commodity hardware.

Providers

Many large healthcare providers maintain millions of records for millions of patients, including semi-structured reports such as radiology images, unstructured doctors' notes, and data captured in spreadsheets and other common computer applications. Also, new models of collaborative care require constant ingestion of new data, integration of massive amounts of data, and updates in near real time to patient records. Data also is being used for predictive analytics for population health management and to help hospitals anticipate and reduce preventable readmissions.

Payers

Many major health insurers support the accountable care organization (ACO) model, which reimburses providers with pay-for-performance, outcome-based-reimbursement incentives. Payers need outcomes data to calculate provider outcomes scores and set reimbursement levels. Also, data management is essential to determine baseline performance and meet Centers for Medicare and Medicaid Services (CMS) requirements for data security, privacy, and HIPAA Safe Harbor guidelines. Additionally, payers are taking advantage of data analytics to predict and minimize claims fraud.

Pharmaceutical industry

R&D for drug development involves enormous volumes of data and many data types, including clinical details, images, labs, and sensor data. Because drug development takes years, any streamlining of processes can pay big dividends. In addition to cost-effective data storage and management, some pharmaceutical companies are using managed data lakes to increase the efficiency of clinical trials, such as speeding up patient recruitment and reducing costs with risk-based monitoring approaches.

Personalized medicine

We're heading in the direction where we'll use data about our DNA, microbiome, nutrition, sleep patterns, and more to customize more effective treatments for disease. A data lake allows for the collection of hundreds of gigabytes of data per person, generated by wearable sensors and other monitoring devices. Integrating this data and developing predictive models requires advanced analytics approaches, making data lakes and self-service data preparation key.

Financial Services

In the financial services industry, managed data lakes can be used to comply with regulatory reporting requirements, detect fraud, more accurately predict financial trends, and improve and personalize the customer experience.

By consolidating multiple enterprise data warehouses into one data lake, financial institutions can move reconciliation, settlement, and regulatory reporting, such as Dodd-Frank, to a single platform. This dramatically reduces the heavy lifting of integration because data is stored in a standard yet flexible format that can accommodate unstructured data.

Retail banking also has important use cases for data lakes. In this field, large institutions need to process thousands of applications for new checking and savings accounts on a daily basis. Bankers that accept these applications consult third-party risk scoring services before opening an account, yet it is common for bank risk analysts to manually override negative recommendations for applicants with poor banking histories. Although these overrides can happen for good reasons (say there are extenuating circumstances for a particular person's application), high-risk accounts tend to be overdrawn and cost banks millions of dollars in losses due to mismanagement or fraud.

By moving to a data lake, banks can store and analyze multiple data streams and help regional managers control account risk in distributed branches. They are able to find out which risk analysts make account decisions that go against risk information by third parties. Creation of a centralized data catalog of the data in the data lake also supports increased access of nontechnical staff such as attorneys, who can quickly perform self-service data analytics. The

net result is better control of fraud. Over time, the accumulation of data in the data lake allows the bank to build algorithms that automatically detect subtle but high-risk patterns that bank risk analysts might have previously failed to identify.

Telecommunications

The telecommunications sector has some unique challenges as revenues continue to decline due to increased competition, commoditization of products and services, and increased resort to the internet in place of more lucrative voice and messaging services. These trends have made data analytics extremely important to telecommunications companies for delivering better services, discovering competitive advantages, adding new revenue streams, and finding efficiencies.

Telecommunications is extremely rich when it comes to subscriber usage data, including which services customers use and where and when they use them. A managed data lake enables telco operators to more effectively take advantage of their data; for example, for new revenue streams. One interesting use case is to monetize the data and sell insights to companies for marketing or other purposes.

Also, customer service can be a strong differentiator in the telecommunications sector. A managed data lake is an excellent solution to support analytics for improving customer experience and delivering more targeted offers such as tiered pricing or customized data packages. Another valuable use case is using a data lake and data analytics to more efficiently guide deployment of new networks, reducing capital investment and operational costs.

Retail

Retailers are challenged to integrate data from many sources, including ecommerce, enterprise resource planning (ERP) and customer relationship management (CRM) systems, social media, customer support, transactional data, market research, emails, supply chain data, call records, and more to create a complete, 360-degree customer view. A more complete customer profile can help retailers to improve customer service, enhance marketing and loyalty programs, and develop new products.

Loyalty programs that track customer information and transactions and use that data to create more targeted and personalized rewards

and experiences can entice customers to not only to shop again, but to spend more or shop more often. A managed data lake can serve as a single repository for all customer data, and support the advanced analytics used to profile customers and optimize a loyalty program.

Personalized offers and recommendations are basic customer expectations today. A managed data lake and self-service data preparation platform for analytics enable retailers to collect nearly real-time or streaming data and use it to deliver personalized customer experiences in stores and online. For example, by capturing web session data (session histories of all users on a page), retailers can provide timely offers based on a customer's web browsing and shopping history.

Another valuable use case for a managed data lake in retail is product development. Big data analytics and data science can help companies expand the adoption of successful products and services by identifying opportunities in underserved geographies or predicting what customers want.

CHAPTER 5
Looking Ahead

Most companies are at the very beginning stages of understanding with respect to optimizing their data storage and analytics platforms. An estimated 70% of the market ignores big data today, and because they use data warehouses, it is tough for them to quickly accommodate business changes. Approximately 20 to 25% of the market stores some of its data in data lakes using scale-out architectures such as Hadoop and Amazon Simple Storage Service (Amazon S3) to more cost-effectively manage big data. However, most of these implementations have turned into data swamps. Data swamps are essentially unmanaged data lakes, so although they still are more cost effective than data warehouses, they are only really useful for some ad hoc exploratory use cases. Finally, 5 to 10% of the market is using managed, governed data lakes, which allows for energized business insights via a scalable, modern data architecture.

As the mainstream adopters and laggards are playing catch up with big data, today's innovators are looking at automation, machine learning, and intelligent data remediation to construct more usable, optimized data lakes. Companies such as Zaloni are working to make this a reality.

As the data lake becomes an important part of next-generation data architectures, we see multiple trends emerging based on different vertical use cases that indicate what the future of data lakes will look like.

Logical Data Lakes

We are seeing more and more requirements for hybrid data stores, in which data can be stored not only in Hadoop Distributed File System (HDFS), but also in object data stores, such as Amazon Simple Storage Service (Amazon S3) or Microsoft Azure Elastic Block storage, or in No-SQL databases. To make this work, enterprises need a unified view of data that exists across these multiple data stores across the multiple environments in an enterprise. The integration of these various technologies and stores within an organization can lead to what is, in effect, a logical data lake. Support for it is going to be critical for many use cases going forward.

Federated Queries

Federated queries go hand-in-hand with logical data lakes. As data is stored in different physical and virtual environments, you might need to use different query tools and decompose a user's query into multiple queries—sending them to on-premises data stores as well as cloud-based data stores, each of which possesses part of the answer. Then, the answers are aggregated and combined and sent back to the user such that they get one version of the truth across the entire logical data lake.

Enterprise Data Marketplaces

Another trend is to make data available to consumers via rich metadata data catalogs in a shopping cart. Many enterprises are already building these portals out of shared Hadoop environments, where users can browse relevant parts of the data lake and have an Amazon-like shopping cart experience in which they select data based on various filters. They can then create a sandbox for that data, perform exploratory ad hoc analytics, and feed the results back into the data lake to be used by others in the organization.

Machine Learning and Intelligent Data Lakes

Not far off in the future are more advanced data environments that use automation and machine learning to create intelligent data lakes. With machine learning, you can build advanced capabilities such as text mining, forecast modeling, data mining, statistical model build-

ing, and predictive analytics. The data lake becomes "responsive" and "self-correcting," with an automated data life cycle process and self-service ingestion and provision. Business users have access and insight into the data they need (for instance 360-degree views of customer profiles), and they don't need IT assistance to extract the data that they want.

The Internet of Things

As the Internet of Things (IoT) continues to grow, much of the data that used to come in via a batch mode is coming in via streaming because the data is being generated at such high velocity. In such cases, enterprises are beginning to keep the data in memory for near-real-time streaming and analytics, to generate insights extremely quickly. This adds another dimension to data lakes—that is, not just being able to process high volumes of data at scale, but to provide low-latency views of that data to the enterprise so that it can react and make better decisions on a near-real-time basis.

In Conclusion

Big data is an extraordinary technology. New types of analysis that weren't feasible on data warehouses are now widespread.

Early data lakes were trumpeted as successes based on their low cost and agility. But as more mainstream use cases emerged, organizations found that they still needed the management and governance controls that dominated in the data warehouse era. The data lake has become a middle ground between data warehouses and "data swamps" in offering systems that are still agile and flexible, but have the safeguards and auditing features that are necessary for business-critical data.

Integrated data lake management solutions like the Zaloni Data Platform (ZDP) are now delivering the necessary controls without making big data as slow and inflexible as its predecessor solutions. Use cases are emerging even in sensitive industries like healthcare, financial services, and retail.

Enterprises are also looking ahead. They see that to be truly valuable, the data lake can't be a silo; rather, it must be one of several platforms in a carefully considered end-to-end modern enterprise data architecture. Just as you must think of metadata from an enterprise-

wide perspective, you need to be able to integrate your data lake with external tools that are part of your enterprise-wide data view. Only then will you be able to build a data lake that is open, extensible, and easy to integrate into your other business-critical platforms.

A Checklist for Success

Are you ready to build a data lake? Following is a checklist of what you need to make sure you are doing so in a controlled yet flexible way.

Business-Benefit Priority List

As you start a data lake project, you need to have a very strong alignment with your business's current and upcoming needs. After all, the data lake needs to provide value that the business is not getting from its data warehouse. This might be from solving pain points or by creating net new revenue streams that you can enable business teams to deliver. Being able to define and articulate this value from a business standpoint, and convince partners to join you on the journey, is very important to your success.

Architectural Oversight

After you have the business alignment and you know what your priorities are, you need to define the upfront architecture: what are the different components you will need, and what will the end technical platform look like? Keep in mind that this is a long-term investment, so you need to think carefully about where the technology is moving. Naturally, you might not have all the answers upfront, so it might be necessary to perform a proof of concept to get some experience and to tune and learn along the way. An especially important aspect of your architectural plans is a good data-management strategy that includes data governance and metadata, and how you will capture that. This is critical if you want to build a managed and governed data lake instead of the much-maligned "data swamp."

Security Strategy

Outline a robust security strategy, especially if your data lake will be a shared platform used by multiple lines of business units or both internal and external stakeholders. Data privacy and security are

critical, especially for sensitive data such as protected health information (PHI) and personally identifiable information (PII). Data that might have been protected before as a result of physical isolation is now available in the data lake. You might have regulatory rules to which you need to conform. You must also think about multitenancy: certain users might not be able to share data with other users. If you are serving multiple external audiences, each customer might have individual data agreements with you, and you need to honor those agreements.

I/O and Memory Model

As part of your technology platform and architecture, you must think about how your data lake will scale out. For example, are you going to decouple the storage and the compute layers? If that's the case, what is the persistent storage layer? Already, enterprises are using Azure or S3 in the cloud to store data persistently, but then spinning up clusters dynamically and spinning them down again when processing is finished. If you plan to perform actions like these, you need to thoroughly understand the throughput requirements during data ingestion, which will also dictate throughput for storage and network as well as whether you can process the data in a timely manner. You need to articulate all this upfront.

Workforce Skillset Evaluation

For any data lake project to be successful, you need to have the right people. You need experts who have previous hands-on experience building data platforms, and who have extensive experience with data management and data governance so that they can define the policies and procedures upfront. You also need data scientists who will be consumers of the platform, and bring them in as stakeholders early in the process of building a data lake to hear their requirements and how they would prefer to interact with the data lake when it is finished.

Operations Plan

Think about your data lake from a service-level agreement (SLA) perspective: what SLA requirements will your business stakeholders expect, especially for business-critical applications that can have impacts on revenues? You need proper SLAs to specify acceptable

downtime, and acceptable quantities of data being ingested, processed, and transformed in a repeatable manner. Going back to the people and skills point, it's critical to have the right people with experience managing these environments, to put together an operations team to support the SLAs and meet the business requirements.

Disaster Recovery Plan

Depending on the business criticality of your data lake as well as the different SLAs you have in place with your different user groups, you need a disaster recovery plan that can support it.

Communications Plan

After you have the data lake platform in place, how will you advertise the fact and bring in additional users? You need to get different business stakeholders interested and show some successes for your data lake environment to flourish because the success of any IT platform ultimately is based upon business adoption.

Five-Year Vision

Given that the data lake is going to be a key foundational platform for the next generation of data technology in enterprises, organizations need to plan ahead on how to incorporate data lakes into their long-term strategies. We see data lakes taking over data warehouses as organizations attempt to be more agile and generate more timely insights from more of their data. Organizations must be aware that data lakes will eventually become hybrids of data stores, include HDFS, NoSQL, and Graph DBs. They will also eventually support real-time data processing and generate streaming analytics—that is, not just rollups of the data in a streaming manner, but machine learning models that produce analytics online as the data is coming in and generate insights in either a supervised or unsupervised manner. Deployment options are going to increase, also, with companies that don't want to go into public clouds building private clouds within their environments, using patterns seen in public clouds.

About the Author

Ben Sharma, CEO and cofounder of Zaloni, is a passionate technologist with experience in solutions architecture and service delivery of big data, analytics, and enterprise infrastructure solutions. Previously with NetApp, Fujitsu, and others, Ben's expertise ranges from business development to production deployment in a wide array of technologies, including Hadoop, HBase, databases, virtualization, and storage. Ben is the coauthor of *Java in Telecommunications* and holds two patents.