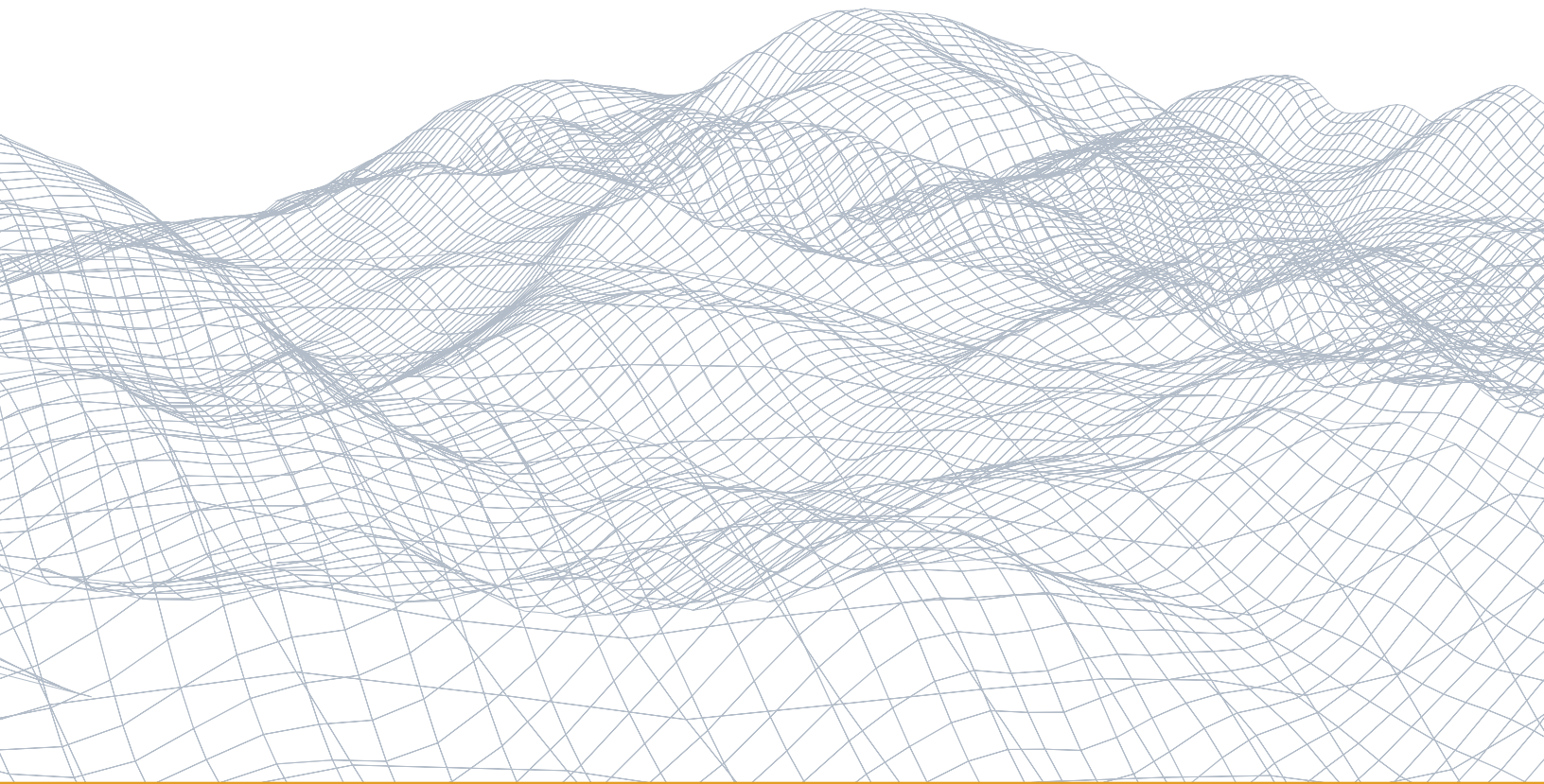


# The Data Lake Reference Architecture

---

*Leveraging a Data Reference Architecture to Ensure Data Lake Success*



# Introduction

Companies are bursting at the seams with data, including from databases and applications, and with streaming data from ecommerce, social media, apps, and connected devices on the Internet of Things (IoT). They are looking for ways to leverage this data to transform their business—gleaning new business insights to create future products and services, revolutionize customer service, streamline operations, uncover new revenue streams and more.

Companies are realizing that traditional technologies can't support their digital transformation and won't enable them to meet their new business needs. As a result, many organizations are modernizing their data platforms, turning to scale-out architectures such as data lakes.

But a poorly architected data lake is also limiting. It can result in a data swamp where data is dumped indiscriminately, providing limited visibility, and thus, limited value. Architecting a data lake for success requires a thoughtful approach. This paper outlines a reference architecture for a data lake that can set you on the right path.

## This paper will discuss:

- Zaloni's data lake reference architecture
- Future-proofing your data lake stack
- Example case studies with data flows

## Data Lake Reference Architecture

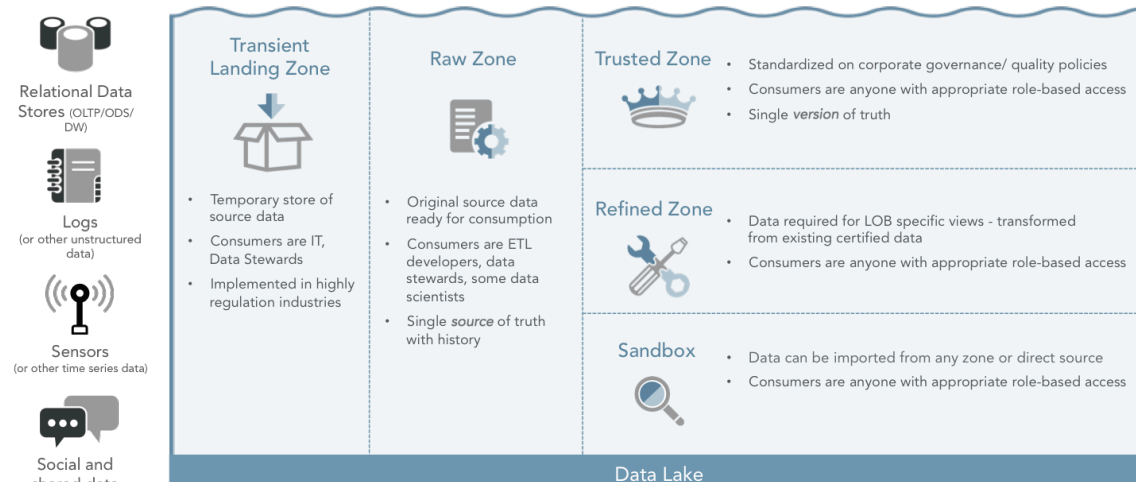
A reference architecture is a framework that can be referred to for 1) understanding industry best practices, 2) tracking a process, 3) providing a template for solutioning, and 4) understanding structures and elements.

Zaloni's data lake reference architecture provides a functional view, and shows how a data lake can be optimally structured and organized to balance flexibility and agility with governance and quality.

A reference architecture is a framework that can be referred to for:

1. Understanding industry best practices
2. Tracking a process
3. Providing a template for solutioning
4. Understanding structures and elements

## Data Lake Reference Architecture



Our reference architecture is organized into four zones, plus a sandbox. Throughout the zones, data is tracked, validated, cataloged, assigned metadata, refined, and more. These capabilities and the zones in which they occur helps us understand what stage the data is in, and what measures have been applied to them thus far.

The key advantage of this architecture is that data can come into the data lake from anywhere, including online transaction processing (OLTP) or operational data store (ODS) systems, a DW, 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.

## 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, where customer information must be encrypted so that it cannot be compromised. In some cases, data must be masked prior to storage.

### Transient Landing Zone

- Temporarily stores source data
- Limited access
- Consumers are IT and Data Stewards
- Implemented in highly regulated industries
- This zone collapses with Raw Zone if security is not needed



The transient zone is temporary, a landing zone for data where security measures can be applied before stored or accessed. With the promise of General Data Protection Regulations (GDPR) being enacted within the next year in the EU, this zone may become even more important as there will be higher levels of regulation and compliance, applicable to more industries.

## Raw Zone

Once the quality checks/security transformations have been performed in the Transient Zone, the data is then loaded into in the Raw Data zone for storage. However, in most situations, a Transient Zone is not needed, and the Raw Zone is the beginning of the data lake journey. Data can be ingested from a wide variety of sources, from the Transient Zone, from relational data stores, from social media feeds, etc.

### Raw Zone

- Original source data
- Ready for consumption
- Treated for basic validation and privacy
- Metadata available to everyone but data access limited based on role
- Consumers are ETL developers, data stewards, some data scientists
- Serves as single "source of truth" with history



Within this zone, data is masked/tokenized as needed, added to catalogs, and metadata is applied. 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.

## Trusted Zone

The trusted zone imports data from the Raw Zone, and 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.”

### Trusted Zone

- Standardized on corporate governance/ quality policies
- Consumers are anyone with appropriate role-based access
- Metadata catalog available to all
- Single *version* of truth



This trusted repository can contain both master data and reference data. Master data is a compilation of the basic data sets that have been cleansed and validated. For example, a healthcare organization may have master data that contain 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 data sets. For example, that healthcare organization might have a reference data set 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.

## Refined Zone

Within the Refined Zone, data is often going 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 lifecycle management. This ensures that the data is in a format from which it can easily be used to create models from. Consumers of this zone are those with appropriate role based access.

### Refined Zone

- Data required for LOB specific views - transformed from existing certified data
- Consumers are anyone with appropriate role-based access
- Metadata catalog available to all




Data is often transformed to reflect the needs of specific LOB's in this zone. For example marketing streams may need to see the ROI of certain engagements to gauge their success, whereas finance departments may need information displayed in the form of balance sheets. Consumers of this zone include those with role-based access, which can be determined by the company.

## Sandbox

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

### Sandbox

- Data can be imported from any zone or directly from source
- Consumers are anyone with appropriate role-based access
- Metadata catalog available to all



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. Some of these insights can be sent directly back to the raw zone, allowing derived data to act as sourced data, and therefore giving data scientists and analysts more with which to work.

## Future-Proofing Your Data Stack

Determining what technologies to employ when building your data lake stack is a complex undertaking. You must consider storage, processing, data management, etc.

In the past, most data lakes resided on-premises. This has undergone a tremendous shift recently, with most companies looking to cloud to replace or augment their implementations. On-premises storage, cloud storage, multi-cloud and hybrid models are all possible with very few tweaks to the corresponding reference architecture. Often the functional architectures look generally the same, while the component architectures are altered to reflect cloud storage platforms and the applications with which they are more compatible.

While on premises storage and processing provide tighter control on data security and data privacy, public cloud systems offer a 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, cloud based-data lakes can also be 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.



## **Data Lake Storage:**

For on-premises data lakes, HDFS seems to be the storage of choice as it provides distributed data with replication. This allows for faster processing of big data use cases. HDFS also allows enterprises to create storage tiers to allow for data life-cycle management leveraging those tiers to save cost while maintaining data retention policies and regulatory requirements.

Cloud-based storage offers a unique advantage as it allows for storage of data decoupled from the need for any compute, thereby allowing enterprises to save on processing costs and leverage different compute powers to meet the use-case demands. Cloud storage also allows for using tiered storage to optimize cost and data retention and regulatory requirements.

## **Data Lake Processing:**

Hadoop has been central to on-premises data lakes as it allows for distributed processing of large data sets across processing clusters for the enterprise. It can also be deployed in a cloud-based data lake to allow for a hybrid data lake using a single distribution (e.g. Cloudera, Hortonworks, and MapR).

Apache Spark provides a faster engine for large-scale data processing leveraging 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 Beam provides an abstraction on top of the processing cluster. By utilizing Beam, enterprises can develop their data processing pipelines using Beam SDK, and then choose a Beam Runner to run the pipeline on a specific large-scale data processing system. The runner can be anything from a Direct Runner, Apex, Flink, Spark, Dataflow, and Gearpump (incubating). This design allows for the processing pipeline to be portable across different runners, thereby provides flexibility to the enterprises to leverage the best platform to meet their data processing requirements in a future-proof way.

## **Data Management:**

The need to manage data across a data lake's varied technology stack magnifies the importance of a unified data management platform. A robust data management platform allows enterprises to manage their data across various storage, compute and processing layers while maintaining clear track of data throughout its lifecycle. In addition, it can automate movement and processing of data within and between zones.

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. Below are four key data management capabilities that must be considered when architecting a next-generation data lake.

### **Metadata Management:**

Cataloging metadata in the data lake is essential. Metadata allows for a way to categorize data and provides context as to when it was uploaded, where it came from, and what lines of business for which it is relevant. This allows one to more easily query data as well as organize it. Metadata serves as the basis for all processing and governance of data across the zones.

### **Security:**

Securing data in the lake is critical. You can leverage role-based access, and can do so within zones. Further, masking and tokenization are two common ways of ensuring that sensitive information remains under wraps until it has been analyzed in the proper way and incorporated into the right models. Masking involves creating a "structurally similar but inauthentic" version of data so that it can be used for testing and training. Tokenization, involves replacing protected data with a value that refers to the data without exposing it.

[www.zaloni.com](http://www.zaloni.com)

## Data Quality:

This refers to ensuring the data is as reflective as possible, and that there is one source of truth that can be referred to by all LOB's, even if initial inputs were more variable.

## Data Lifecycle Management:

Management of the data lifecycle involves cold, warm, and hot designations, in which one tracks how long data has been in the data lake for, and how long it should continue to remain for before being recycled.

To address these capabilities across the data lake companies consider a data management platform such as the Zaloni Data Management Platform (ZDP).

## Case Studies with Data Flows

Companies are making serious strides modernizing their data architectures with data lakes, addressing old problems in new ways and creating new opportunities to enhance their business, their customer loyalty and their competitive advantage. Below we outline 2 distinct case studies that leveraged a reference architecture to build a clean, actionable data lake and were well-rewarded for the effort.

One of these implementations was on-premises and the other leveraged a cloud-first strategy on AWS. For each of these examples, we will show a data flow to explain what tools are required to accomplish a particular use case.

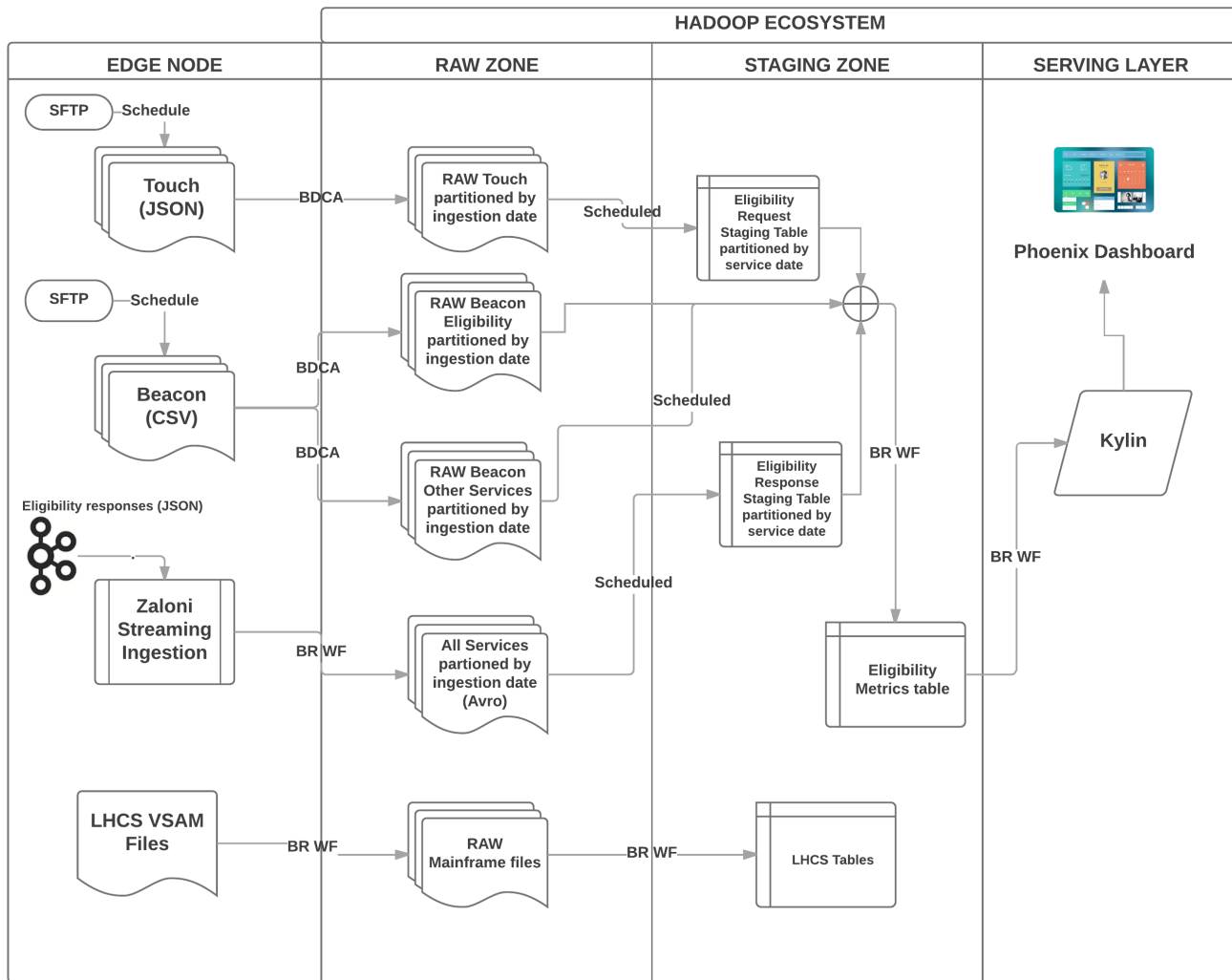
In these data flows, processes such as ingestion, storage, operations, and security are all addressed, as well as different subsections, such as batch processing or dashboard operations displays. For each of these subsections, recommendations are made and technology is stitched together to ensure a seamless implementation.

### On-premises data lake for healthcare diagnostics

One of the largest healthcare diagnostics companies in the world, had a large mainframe environment to handle the high volumes of patient data used in reports for thousands of clients. The environment was inflexible, preventing the company from being able to scale to add new clients or increase computing capacity to speed up reporting to meet existing client demand for more frequent reports. The customer also was unable to generate custom reports for specific requests; for example, a report for a specific patient. The customer needed a new infrastructure to automate processes and scale processing capacity as required.

**With the data lake:** Zaloni took a multi-step approach to the project. First the team built a data lake and put it into production, leveraging ZDP for ingestion, cataloging, data management and governance. New client data was ingested into the data lake, and custom code was developed to generate reports according to clients' preferences. The customer has now begun to add existing clients to the new platform, ingesting mainframe files/historical data into Hadoop.

**Reference Architecture Flow:** The data lake is operating within a larger Hadoop ecosystem with capabilities ranging from ingestion to both batch and stream processing. Choosing an on-premises implementation, the customer utilized the java based HDFS system as its distributed file system and a form of data storage, as well as a multitude of other programs.



**Results:** With significantly higher reporting flexibility and agility, the customer is able to meet client demand for daily reporting. The company has realized reduced data processing, storage, licensing and maintenance costs, and will have nearly unlimited capacity to add new clients into the future.

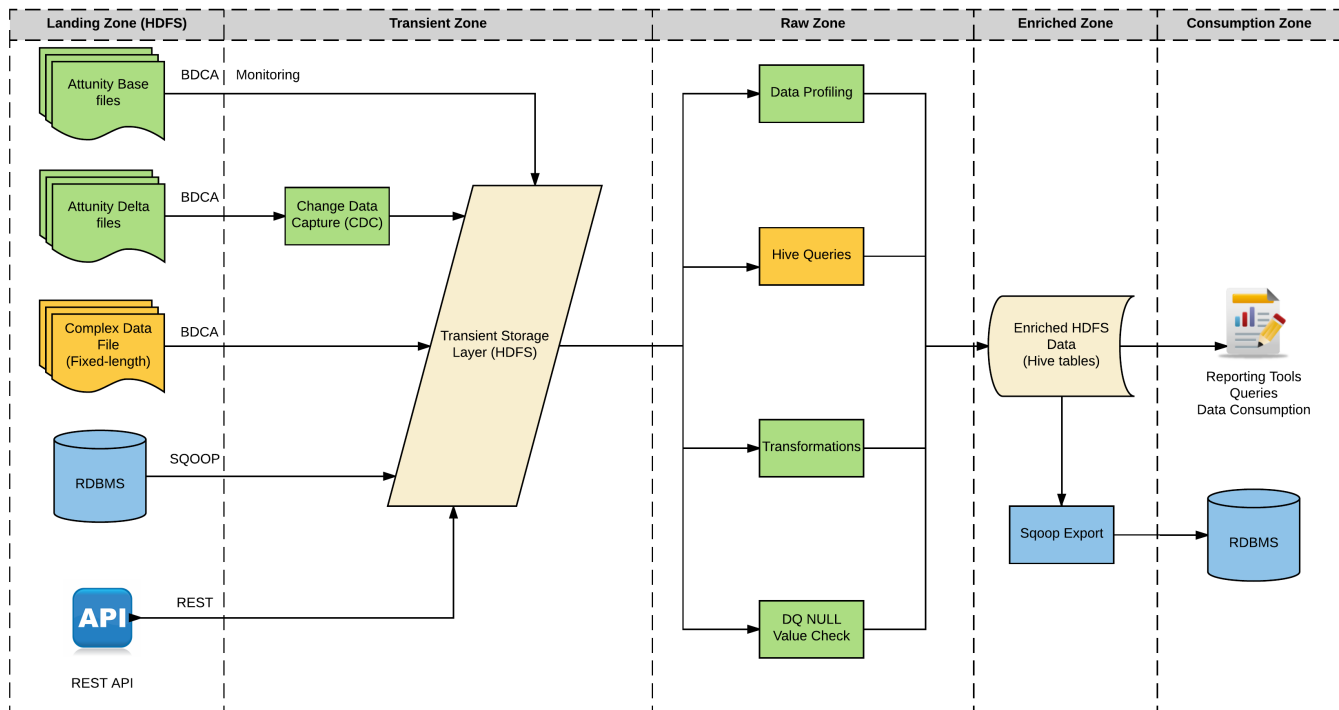
### Cloud data lake for Customer 360 initiative in publishing company

One of the largest publishers in North America wanted to augment its business model by monetizing its 45 years of subscriber data to enable a Customer 360 approach and deliver marketing insights to its customers. The company had decades of valuable subscriber data from its magazines, online subscriptions and marketing materials, untapped and sitting in siloed systems across the organization. It was unable to combine these large volumes of internal data with third-party data to create additional downstream revenue sources.

**With the data lake:** A new cloud-based data lake provided a centralized repository for internal and third-party data. In support of the company's Customer 360 strategy, the data lake enabled the company to gather data from all of its systems and manage and govern it with a unified data platform, ZDP.



**Reference Architecture Flow:** The company chose to host its data lake in the cloud, utilizing Amazon's Simple Storage Service as a data store and Redshift as a relational data store. The company chose to implement a transient zone such that data could be protected and qualified before being permanently stored.



**Results:** The company has achieved four times the functionality in half the time of building in-house. The customer expects to realize significant savings due to reduced time to market and the ability to seamlessly scale data management and analytics at the speed of the business.

## Architecting Your Data Lake for Success

Data lakes have a host of abilities that can serve a wide variety of use cases and industries across multiple lines of business. However, to extract maximum value from the data lake, it is crucial to ensure that they are properly architected and managed.

Using a reference architecture is a good way to structure a data lake, as well as plan out its implementation and functions. The Zaloni data lake reference architecture serves as a good functional model from which to draw.

As you architect your data lake, regardless of deployment model, remember to address data management and governance. We suggest an integrated platform, such as Zaloni Data Management Platform (ZDP) in order to build a data lake that can meet your business needs today and that can scale with you in the future.

# ZALONI DATA PLATFORM

The Zaloni Data Platform (ZDP) is an award-winning self-service data platform providing the capabilities required for data management, governance and self-service to deliver a production-ready data lake that can be deployed in hybrid, cloud or multi-cloud environments. The platform operationalizes data management and eliminates data silos for central management of all enterprise data sources, regardless of location and delivers a self-service, enterprise-wide data catalog through which to discover and wrangle data sets, and derive transformational business insights using advanced analytics.

## Simplify and amplify your modern data infrastructure

Our fully integrated platform minimizes the pain of modernizing your data infrastructure and helps you streamline your complex and fragmented data stack:

- Leverages any store, any distribution, any deployment - cloud, multi-cloud, on-premises or hybrid
- Integrates easily into your existing data architecture
- Simplifies use, management and integration of ever-changing big data technologies and tools

## Platform Benefits

The benefits of ZDP are significant. Across Zaloni's many customer implementations, it has been found that Zaloni makes it 75% faster to deploy a data lake. Additionally, third-party research has shown that leveraging Zaloni's platform can result in a 650% return on investment versus implementing a DIY solution of legacy and homegrown tools.

### SCALE

Accelerate time and effort to build the data lake by **75%**.



### SPEED

Reduce time to insight from week to **minutes**.



### VALUE

Enjoy a **650%** return on investment vs. build your own.



## About Zaloni

### Operationalize your data lake. Accelerate business insights.

Zaloni simplifies big data for transformative business insights. We work with pioneering enterprises to modernize their data architecture and operationalize their data lakes to incorporate data into everyday business practices. The Zaloni Data Platform (ZDP) provides total control throughout the data pipeline from ingestion to analytics, with comprehensive data management, governance and self-service data preparation capabilities for IT and business users. A leader in big data for more than a decade, Zaloni's expertise is deep, spans multiple industries, and has proven invaluable to customers at many of the world's top companies. We are proud to be recognized by CRN's 2018 Big Data 100 list, Forbes top 20 big data companies to work for, and Red Herring's Top 100 North America Award.

#### To learn more:

Call us: +1 919.323.4050

E-mail: [info@zaloni.com](mailto:info@zaloni.com)

Visit: [www.zaloni.com](http://www.zaloni.com)

