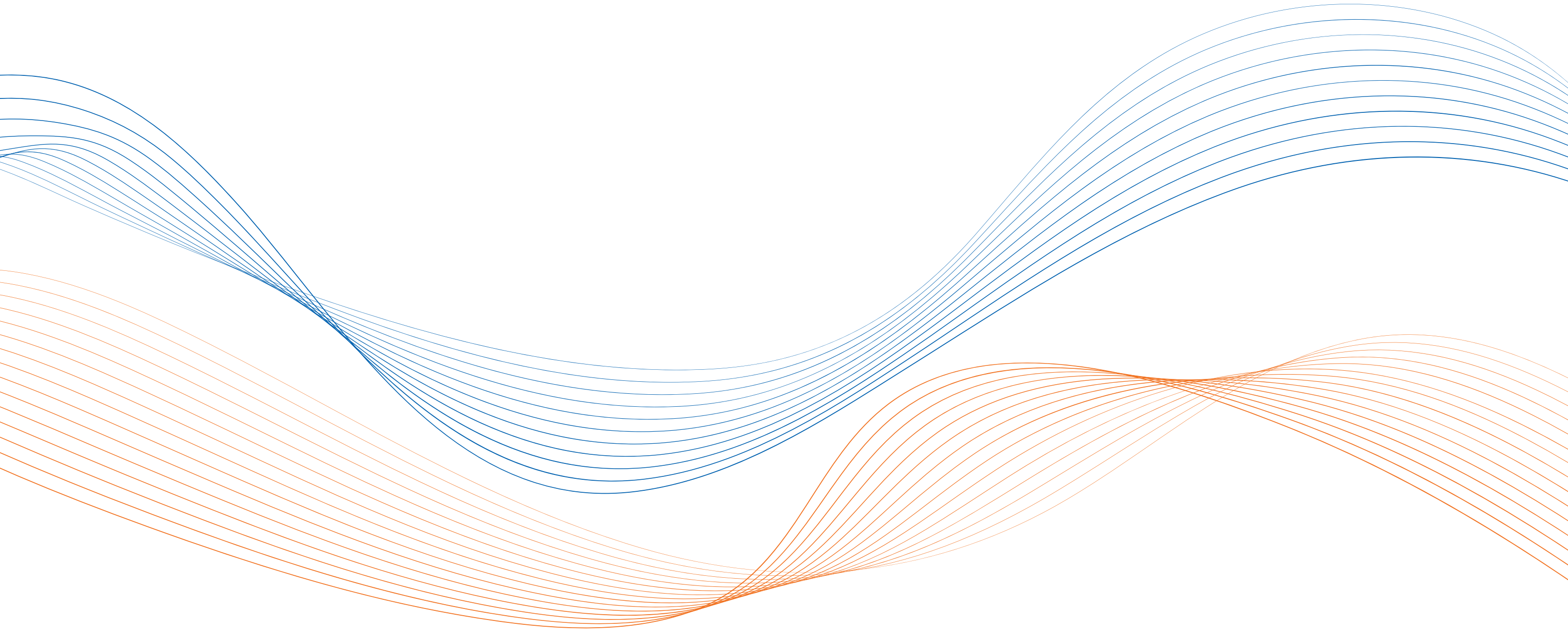
EndZone Governance™ Architecture

How Zaloni's EndZone Governance™ zone-based architecture provides security and control for DataOps success



zaloni

www.zaloni.com

Introduction

Companies are bursting at the seams with data, including data from various databases and applications, streaming data from e-commerce, social media and mobile apps, and connected devices through 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.

As data increases, so does the complexity. Companies today are challenged with data sprawl, mixed cloud and on-premises environments and emerging 3rd party data sources. Additionally, because of experiences with the COVID-19 pandemic, companies want data agility and flexibility to quickly adapt during times of disruption and be better prepared for future crises.

Companies are realizing that traditional technologies can't support their data platform modernization and the DataOps approach, hindering their ability to meet their new business needs. As a result, many companies are modernizing their data platforms, and turning to DataOps for visibility and control of the data "supply chain"

The key to a successful modern data platform is a data architecture that provides data governance and visibility at every step in the data supply chain.

This paper outlines Zaloni's unique EndZone Governance™ reference architecture that has been implemented for many of today's most innovative companies.

This paper will discuss:

- Zaloni's EndZone Governance architecture
- Future-proofing your data technology stack
- Example case studies with data flows

Benefits of EndZone Governance:

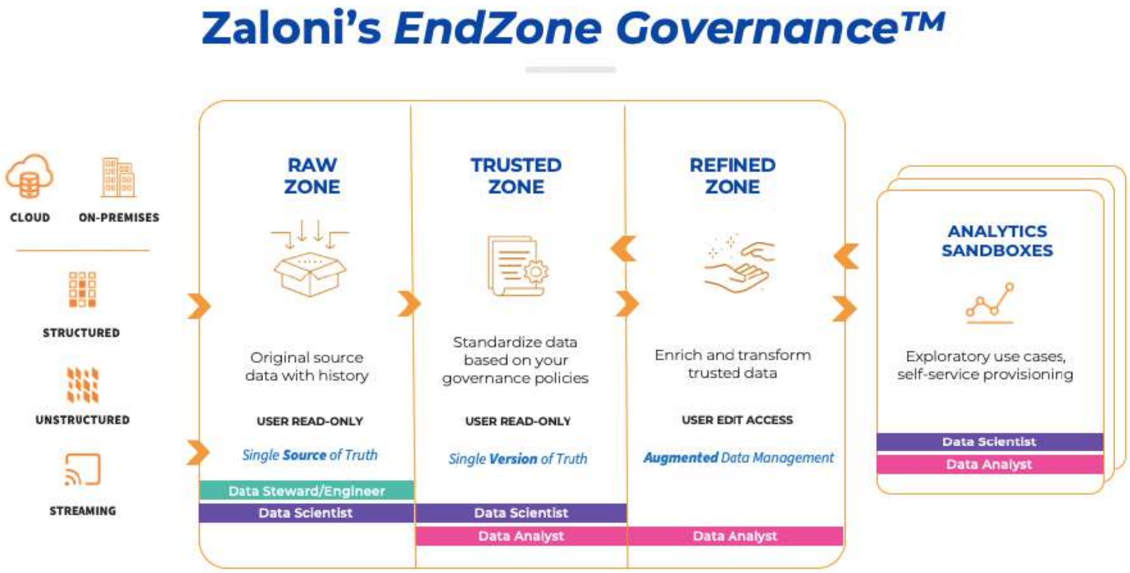
- Trust that the right people have access to the right data at every step of the supply chain
- Increase data quality with automated controls to eliminate manual errors
- Confidence knowing data is protected and secured

EndZone Governance™ 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 data structures and elements.

Zaloni's EndZone Governance™ architecture is reference architecture based on hundreds of data platform implementations. This architecture leverages a zone-based approach through which data can live and travel throughout its lifecycle.

Zaloni's EndZone Governance™ architecture provides a functional view that shows how a data environment can be optimally structured and organized to balance flexibility and agility with governance and quality.



Our reference architecture is organized into three zones, plus analytics sandboxes. 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 be ingested or cataloged from anywhere, including online transaction processing (OLTP) or operational data store (ODS) systems, data warehouses, logs or other machine data, or from cloud sources. These source systems include many different formats, such as file data, database data, ETL, streaming data, and even data coming through APIs.

Raw Zone

Data is ingested from a wide variety of sources including relational data stores, social media feeds, etc. 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."

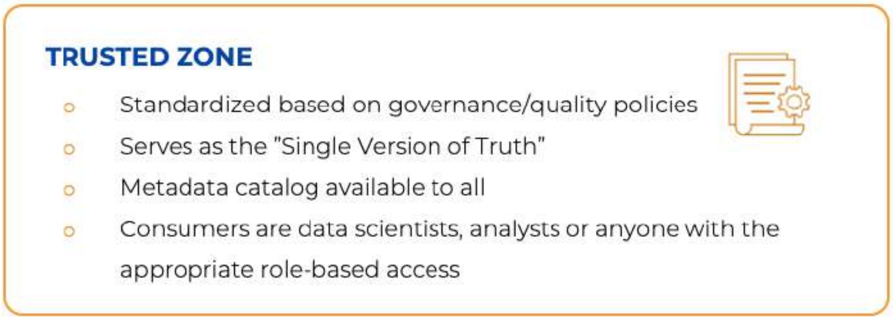
RAW ZONE

- Original source data with history
- Treated for basic validation
- Metadata available to everyone
- Consumers are engineers, data stewards and data scientists
- Serves as the "Single Source of Truth"

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

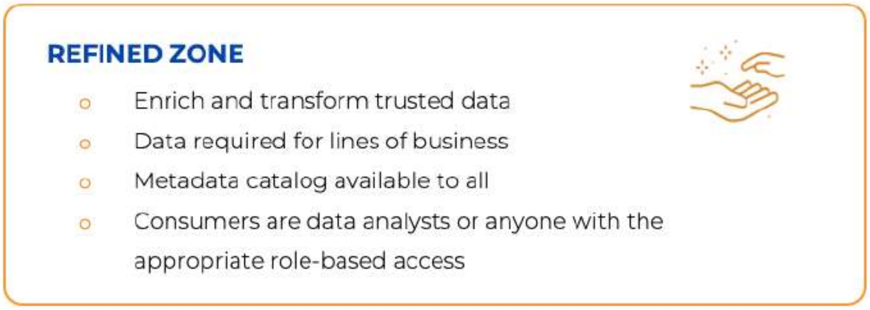
- Standardized based on governance/quality policies
- Serves as the "Single Version of Truth"
- Metadata catalog available to all
- Consumers are data scientists, analysts or anyone with the appropriate role-based access

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, a 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 for analytics. 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 analytic models. Consumers of this zone are typically data analysts or those with appropriate role based access.



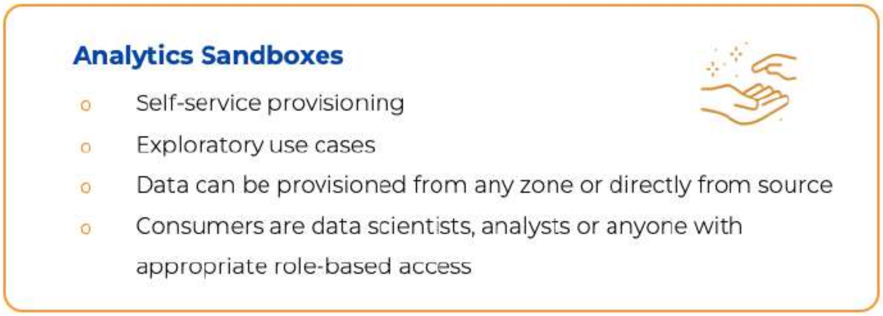
REFINED ZONE

- Enrich and transform trusted data
- Data required for lines of business
- Metadata catalog available to all
- Consumers are data analysts or anyone with the appropriate role-based access

Data is often transformed to reflect the needs of specific lines of business (LOB) 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 to be displayed in the form of balance sheets.

Analytics Sandboxes

An analytics sandbox is integral to a data platform, as it allows data scientists and analysts to create ad hoc exploratory use cases in built analytics environments without having to involve the IT department or dedicate funds to creating suitable environments within which to test the data.



Analytics Sandboxes

- Self-service provisioning
- Exploratory use cases
- Data can be provisioned from any zone or directly from source
- Consumers are data scientists, analysts or anyone with appropriate role-based access

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 stack is a complex undertaking. You must consider storage, processing, data management, etc. Additionally, to successfully leverage a DataOps approach, it's critical that your data stack is tightly integrated to enable end-to-end visibility and control over your data supply chain.

In the past, most data resided on-premises. This has undergone a tremendous shift, 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 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.

On-premises storage and processing provide tighter controls on data security and data privacy. Public cloud systems offer a highly scalable 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. Additionally, with the rapidly changing tools and technologies in the ecosystem, cloud environments 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.

Storage:

For on-premises environments, 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 lifecycle 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.

Processing:

Hadoop has been central to on-premises data environments 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 environment using a single Hadoop distribution.

Apache Spark provides a faster engine for large-scale data processing leveraging in-memory computing. It can run on Hadoop, Mesos, in the 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:

With DataOps, the need for agility and extensibility to manage data across a company's varied technology stack magnifies the importance of unified data management. A robust data management platform allows enterprises to manage their data across various storage, compute and processing layers while maintaining clear tracking of data at every stage of the data supply chain. 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 three key data management capabilities that must be considered for a successful DataOps environment:

- **Metadata Management:** Cataloging metadata across the organization 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 and Access Control:** Securing data is critical for proper data governance. Using role-based access controls you can decide who has access to each zone. Further, masking and tokenization are two common ways of ensuring that sensitive information remains protected 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.

- **Collaborative Data Catalog:** Collaborative data catalogs provide data analysts and scientists with quick self-service access to trusted data. For successful DataOps, the catalog should be more than just an inventory of data, it should provide self-service preparation and provisioning capabilities along with collaboration features to tag, annotate and share data across teams to increase productivity and improve cross-team insights.

To address these capabilities companies consider an end-to-end data platform such as Zaloni's Arena platform.

Case Studies with Data Flows

Companies are making serious strides modernizing their data architectures, 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 environment and were well-rewarded for the effort.

One of these implementations was in the cloud on AWS and the other on-premises. 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.

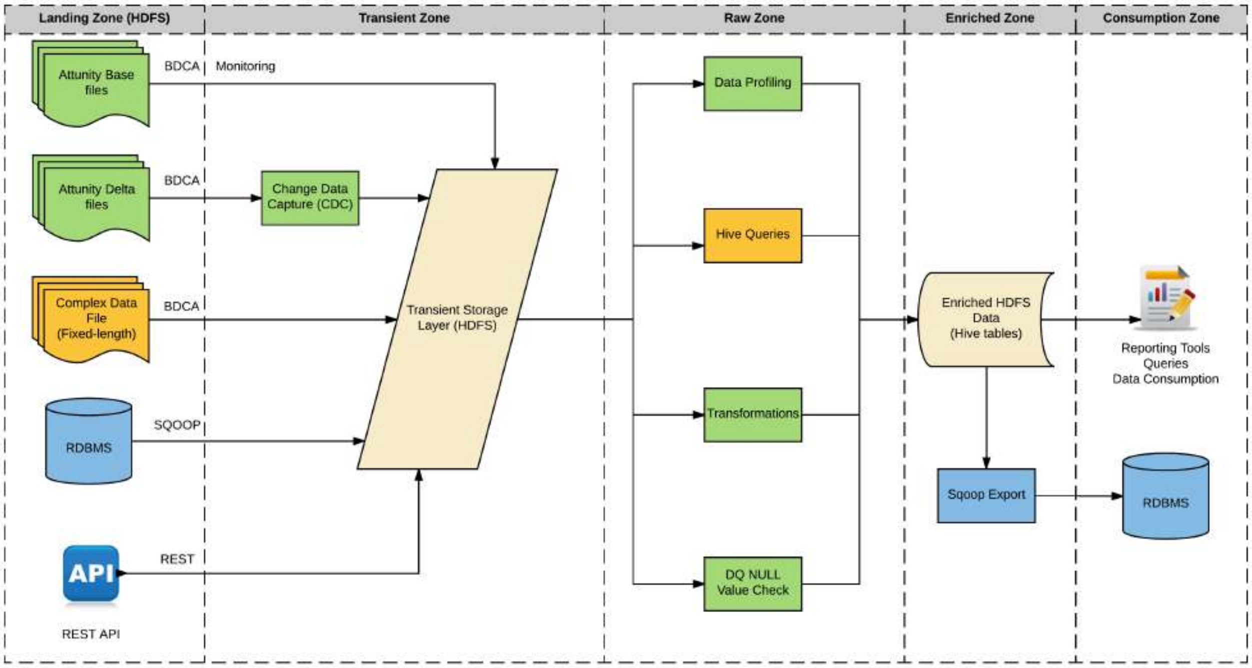
Cloud data lake for Customer 360 initiative in publishing company

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.

Challenge: The company was limited by existing hardware and software and was unable to combine the large volumes of internal data with third-party data to provide its customers with access to subscriber data – missing an opportunity to create additional downstream revenue sources. The company began to go down the path of custom open source development, but found this effort was consuming too much time and too many resources to be sustainable. They knew they wanted to leverage the cloud for its cost efficiency and scalability but needed a way to manage and govern their data in the cloud.

Solution: The company used Arena on AWS to build a managed and governed cloud-based data lake that provided a scalable, centralized repository for all 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, track lineage and govern consistently across the enterprise with a single unified data platform. Arena enabled a stable, reliable foundation on which the company could implement and customize an operational framework according to their data needs. Arena also provided an intuitive, user-friendly system that required minimal customization but was fully extensible to meet customer requirements.

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: Throughout implementation, the company achieved 4 times the functionality in 1/2 the time of building in-house. The company expects to realize significant savings due to faster ramp up to using big data technologies, reduced time to market and the ability to seamlessly scale data management and analytics at the speed of the business.

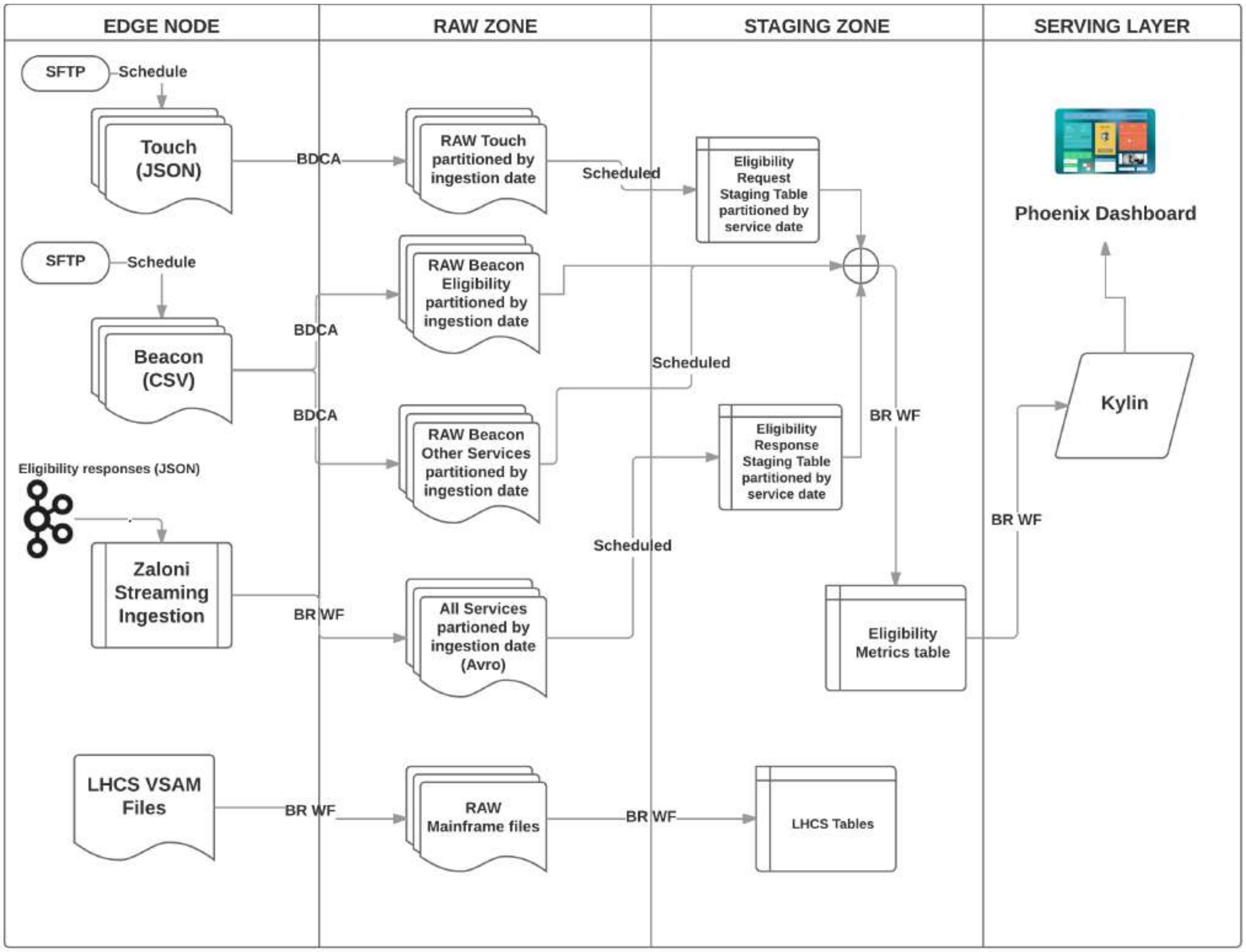
Centralized data lake enables revenue-cycle analytics

One of the largest healthcare diagnostics companies in the world, employs more than 50,000 staff worldwide and serves 220,000 clients, including doctor's offices, hospitals, managed care organizations and biotechnology and pharmaceutical companies, with leading-edge medical laboratory tests and services. The company wanted to improve billings and claims management and visibility into the health of its revenue cycle.

Challenge: The company had a large mainframe environment to handle enormous volumes of customer billing data. Data was stored in silos, making it difficult to do the data correlation needed to produce billing and claims reports. It was too onerous and time consuming for the business to ask simple questions about billing and claims; therefore, reports were only generated on a monthly basis. The company wanted a faster way to process billing data in order to have a more real-time picture of its business.

Solution: Zaloni offloaded data processing into a data lake and added an OLAP (online analytical processing) layer, which provided a user-friendly front-end foundation for analytics on various datasets. New client data was ingested into the data lake, and custom code was developed to generate reports according to clients' preferences. Arena provided complete end-to-end management and governance of the data lake. Arena's data catalog provided easy self-service access to business users so they can easily find, prepare and provision data to their preferred business intelligence (BI) tools without going through IT.

Reference Architecture Flow: The data lake is operating within a larger data 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.


EDGE NODE
RAW ZONE
STAGING ZONE
SERVING LAYER
SFTP
Schedule
Touch (JSON)
BDCA
RAW Touch partitioned by ingestion date
Scheduled
Eligibility Request Staging Table partitioned by service date
Phoenix Dashboard
SFTP
Schedule
RAW Beacon Eligibility partitioned by ingestion date
BDCA
Beacon (CSV)
Scheduled
BDCA
RAW Beacon Other Services partitioned by ingestion date
BR WF
Eligibility Response Staging Table partitioned by service date
Kylin
Eligibility responses (JSON)
BR WF
Scheduled
Zaloni Streaming Ingestion
BR WF
All Services partioned by ingestion date (Avro)
Eligibility Metrics table
LHCS VSAM Files
BR WF
RAW Mainframe files
BR WF
LHCS Tables

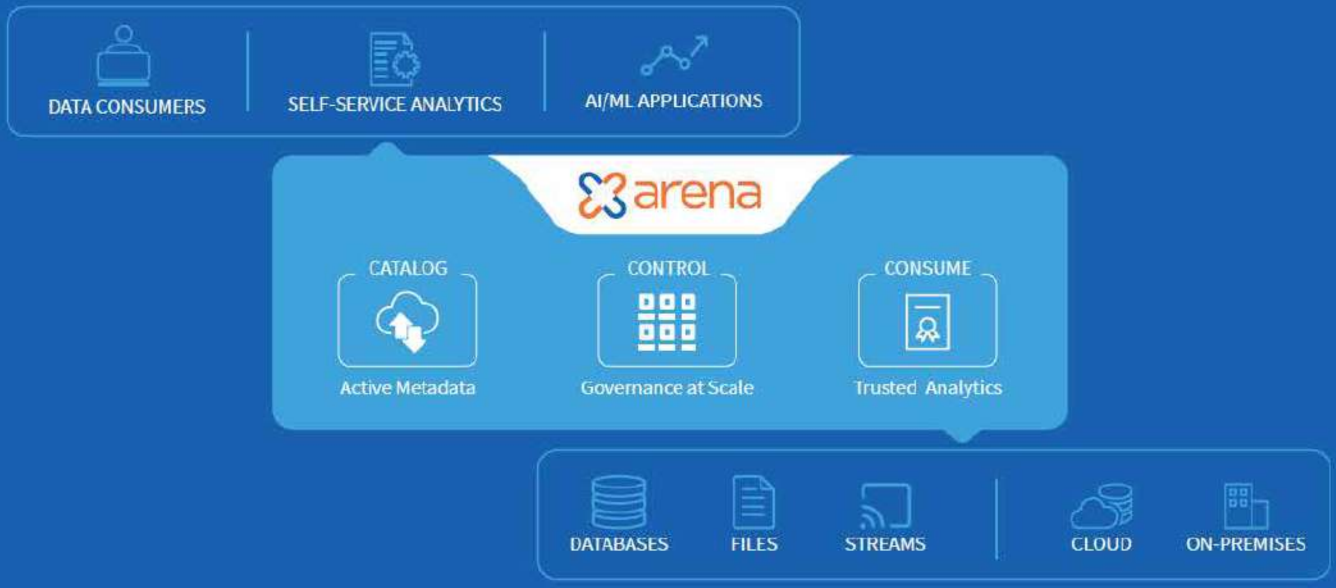
Results: The company has seen multiple benefits from having a unified, historical view of billing data with third party, patient and client-billed data formats in a single data lake platform. With significantly higher reporting flexibility and agility, Business users are able to report billing and claims status daily versus monthly, giving the company increased visibility into revenue cycle KPIs. Business users also are able to customize reports without needing to involve developers from the IT team. Additionally, 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.

EndZone Governance™ and Arena™ for DataOps Success

Modern data environments 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, it is crucial to ensure that they are properly architected and managed.

Using a reference architecture is a good way to structure your data environment, as well as plan out its implementation and functions. Zaloni's EndZone Governance™ architecture serves as a good functional model from which to draw. As you build your data architecture, regardless of deployment model, remember to address data management and governance.

We suggest an unified DataOps platform, such as Arena, that can meet your business needs today and that can scale with you in the future. Visit www.zaloni.com/demo to set up a custom demo and see how Arena™ and EndZone Governance™ can solve your data challenges.



Learn more
About Zaloni

+1 919.323.4050
info@zaloni.com

About Zaloni

At Zaloni, we believe in the unrealized power of data. Our data management software provides an augmented catalog that enables self-service data enrichment and consumption. We work with the world's leading companies, delivering exceptional data governance built on an extensible, machine-learning platform that both improves and safeguards enterprises' data assets.