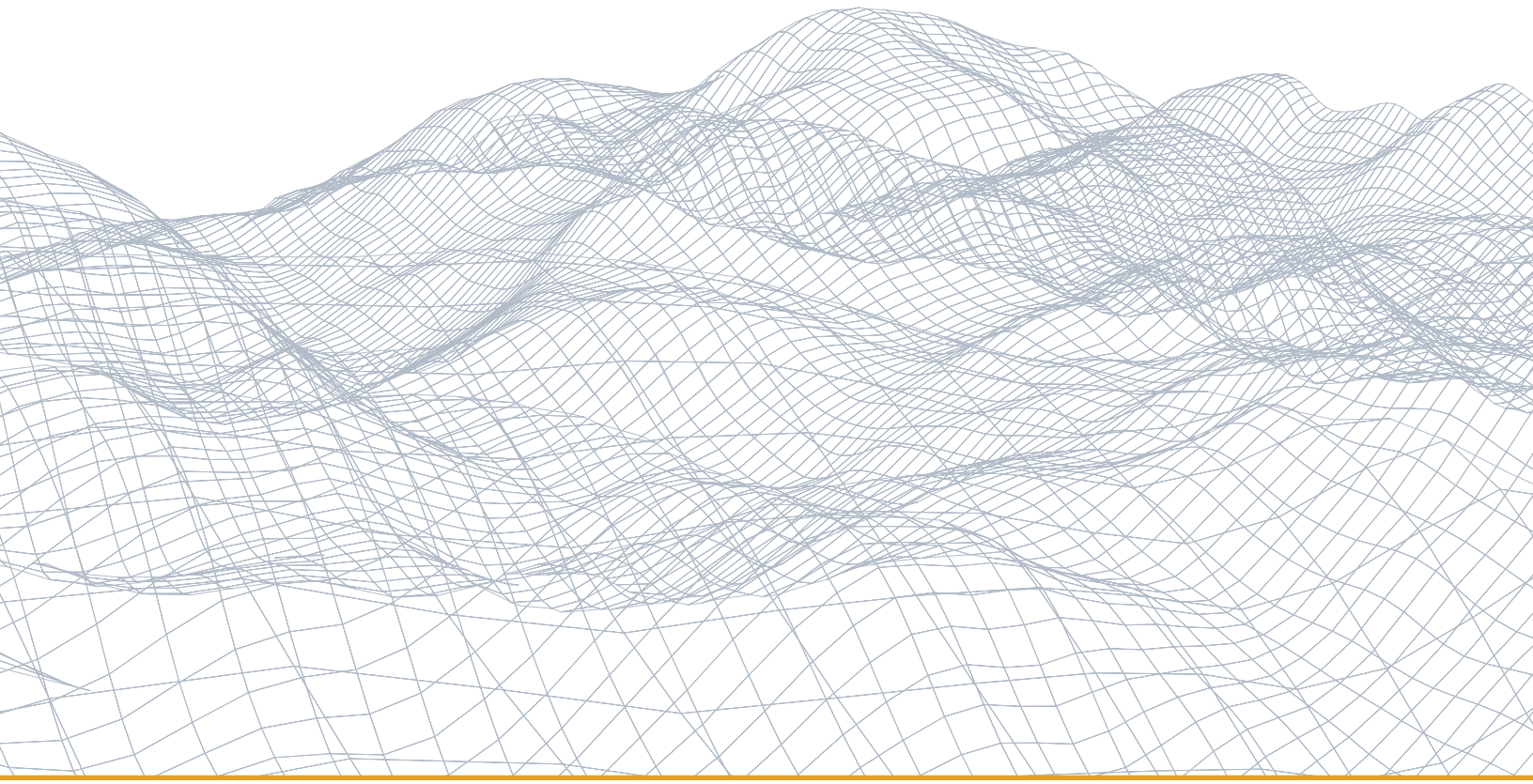
Data Lakes: The 360-Degree Approach

What enterprises need to consider to derive real value from a data lake



ZALONI
THE DATA LAKE COMPANY

www.zaloni.com

Introduction

Enterprises today generate and have access to huge volumes of data from a multitude of sources. Whereas businesses used to consider their data primarily a cost, requiring funding for ever-increasing amounts of storage, now most enterprises consider their data an asset—and are looking for new ways to leverage it for competitive advantage or to improve the bottom line.

It's our position that managed data lakes—centralized repositories for raw data from multiple sources that can be made available to many users for nearly any purpose—will become essential to the modern data architecture. Managed data lakes will be fed by various structured data sources, real-time data streams, such as from the Internet of Things, and unstructured data like emails, videos, photos, audio files, presentations and more.

All of the data will be stored in this centralized repository—whether in the cloud or on premises or a hybrid—where it can be transformed, cleaned and manipulated by data scientists and business users. Then, prepared datasets can be fed back into a traditional enterprise data warehouse for business intelligence analysis, or to other visualization tools for data science, data discovery, analytics, predictive modeling and reporting.

Although enterprises have been using data for business intelligence and marketing for decades, the volume, types and real-time availability of data that is generated today, as well as modern data architectures and tools that can accommodate and analyze all of this data, are changing the ways enterprises can derive value from data. It is the inherent and essential flexibility of data lakes that promises to give enterprises the agility and scalability they require to discover timely, valuable business insights from big data.

The promise of the managed data lake is to allow more cost-effective storage, access and management of all the data in an enterprise, with a goal of providing a comprehensive and governed view of data within the organization.

This paper will discuss:

- The benefits of a well-managed data lake
- The Data Lake 360 approach to delivering on big data
- The most common use cases for data lakes today
- Example customers who have successfully implemented data lake 360

Benefits of a data lake approach

The benefits of integrating a data lake into your overall data architecture can be significant.

1. Cost-savings

Save millions in storage costs and data processing

Scale-out architectures (e.g., Hadoop, S3) can store raw data in any format at a fraction of the cost of a traditional enterprise data warehouse (EDW). In fact, we helped one client achieve 20 times the storage capacity of their EDW at 50% of the cost of a previously planned EDW upgrade. Another client achieved a 100x cost

reduction per terabyte of stored data. This can be particularly beneficial for companies in industries that require long-term data retention, such as financial services and healthcare. Building a data lake in the cloud can reduce storage costs even more, as storage and compute services can be decoupled and paid for at different rates.

In addition to storage, scale-out architectures enable faster loading of data and parallel processing, resulting in faster time to insight. For example, one of our clients quadrupled the throughput of their system after migrating processing to Hadoop. Hadoop is also much more effective than the EDW for processing the increasing amount of unstructured and semi-structured data that's important for analytics today. Furthermore, building a data lake in the cloud enables businesses to spin clusters up and down on demand, eliminating wasted resources and lowering costs.

2. Increased revenue

Drive incremental sales and create new revenue streams

There's a drive to monetize enterprise data—to use data to drive incremental sales or create new revenue streams. Better understanding the preferences and habits of existing customers (a 360 customer view) helps enterprises tailor the customer experience to deliver what customers want. For example, we helped one global mobile and Internet provider capture incremental revenue through offers of tiered pricing or data packages based on visibility into subscriber usage at the individual level. We also helped a large telecom operator use its subscriber geolocation data to identify and create new products and services.

3. Reduced risk

Ensure compliance and prevent fraud

When an enterprise is unable to harness all of its data—which is often scattered across an organization in disjointed systems and databases—to get a complete and timely snapshot of its business, it can leave the door wide open for increased risk. One of our clients looked to close this gap with an automated fraud detection system. We helped this large insurance company by building a platform that ingested more than 4 terabytes of data from multiple systems, crawled public websites and searched healthcare providers in order to stay ahead of fraudsters.

Business benefits of a managed data lake:

- Cost savings in both storage and data processing
- Increased revenue through new data-driven business models
- Risk reduction in terms of regulatory compliance and fraud prevention

Simplifying data lake implementation into 3 phases

Transitioning to a modern data architecture to enable advanced analytics and data science is a complicated process, but a worthwhile one when you consider the long-term benefits of being a data-driven business. Let's simplify a successful data lake implementation into three phases.

1. **Enable the lake:** Build the lake and determine how you will ingest, organize and catalog your data.
2. **Govern the data:** This involves data quality rules, automation workflows, as well as data security.
3. **Engage the business:** Deliver the data to more end users, including business end users, to maximize its value—"democratizing" access to your data. This involves implementing tools that make data discovery, enrichment and provisioning very intuitive for less-technically savvy business users.

Implementing an actionable data lake requires 3 phases:

1. Enable the Lake
2. Govern the Data
3. Engage the Business

Building a managed data lake is a complex undertaking. Some common challenges you may encounter include the following:

Transitioning from POC to production. Your proof of concept (POC) may have gone well, but now you need to operationalize your data lake for new use cases and integrate it into daily business practices across the enterprise.

Navigating the complexity. You have to contend with an ecosystem that includes hardware, software and applications. Hadoop requires you to integrate multiple tools to build a successful managed data lake.

Solving the skills gap. Implementing a managed data lake requires a specific skill set—one that many development and architecture professionals may not have, making talent hard to find and costly to hire. A Gartner survey revealed that 57% of enterprises say they are not ready to adopt Hadoop because of skills gaps.[1]

Keeping up with the big data ecosystem. Big data is a relatively new technology and its ecosystem is constantly changing as the community develops new tools and solutions to increase data availability, and make data processing and analysis faster, storage more efficient and programming simpler.

Remembering your data management fundamentals. Without a systematic and automated way to manage and govern data, you won't know what data you have, be able to trust your data quality, provide access to data for multiple users, or comply with security and privacy regulations.

Challenges when implementing a data lake:

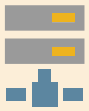
- Transitioning successfully from POC to production
- Navigating the complexity
- Solving the skills gap
- Keeping up with the big data ecosystem
- Remembering your data management fundamentals

How to find data lake success: The 360-degree approach

Many early adopters used data lakes as a relatively inexpensive storage solution and dumped data into them without much of a plan. Now these enterprises are struggling to derive value from the data in the data lake and are unable to operationalize their data lake beyond the original proof of concept (POC). By taking a holistic view of how a managed data lake fits into your overall data architecture, you can ensure that your data is managed and governed properly to make it accessible for data scientists and business users to derive value from it. In order to achieve the agility, shorter time to insight, and scalability that a data lake promises, you need a unified, integrated approach to data lake management and governance that will provide:

Data Lake 360°: A holistic approach to implementing a data lake, ensuring that your data is managed and governed properly to make it accessible and actionable to the business.

Data visibility
Metadata management capabilities allow you to keep track of what data is in the lake and its source, format and lineage.



Data reliability
Gives you confidence that your analytics are always running on the right data, with the right quality.

[1] Gartner Survey Highlights Challenges to Hadoop Adoption. May 13, 2015. http://www.gartner.com/newsroom/id/3051717.

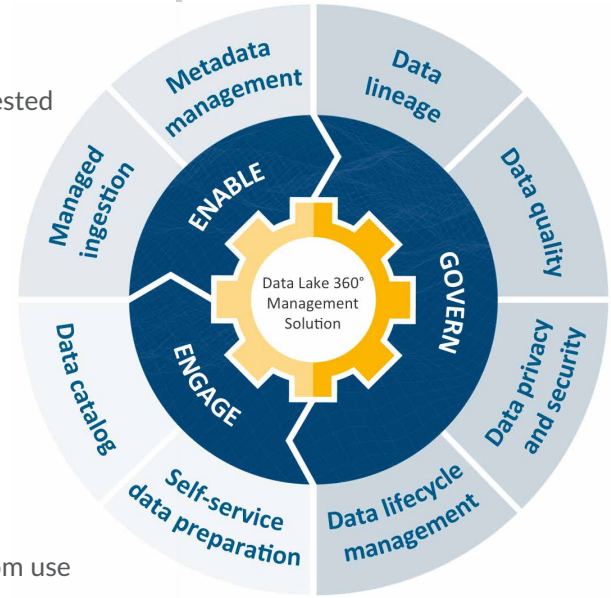
Data security and privacy
Ensures access control, provides data masking (e.g., for personally identifiable information (PII)) and ensures compliance with industry and other regulations.

Democratized access to useful data
Extends end user accessibility and self-service (to those with permissions) to get more value from the data.

Zaloni's Data Lake 360° provides:

- Data visibility
- Data reliability
- Data security and privacy
- Enterprise-wide access to useful data

At Zaloni, our 360-degree approach considers everything you need for holistic management and governance of your data lake, and is comprised of the following capabilities:



Enable the lake

- **Managed ingestion:** Define, track and log all steps of what data is ingested into the data lake
- **Metadata management:** Apply technical, operational and business metadata to have a more complete view of your inventory

Govern the data

- **Data lineage:** Track where data comes from and what happens to it over time
- **Data privacy and security:** Leverage metadata to apply permissions, policy-based security, watermarking, masking and tokenization
- **Data quality:** Provide a clear understanding of the level of each dataset's quality
- **Data lifecycle management:** Automate management of data assets from use and reuse to eventual retirement and long-term storage/archiving

Engage the business

- **Data catalog:** Provide an enterprise-wide view of data to discover and curate
- **Self-service data preparation:** Provide a sandbox for end users to interact and work with sample data and export results or create a workflow

Operationalizing the end-to-end data lake management pipeline

Data Lake 360 is a holistic approach to understanding the required components of a clean, actionable, and scalable data lake. But how do you put it all together? How do you make it real?

You need a data lake management platform that automates and integrates data management. You need a platform that allows you to address the following questions:

- How do you effectively hydrate your data lake?
- How do you know what is in the lake, where it came from, and where it lives in the lake?
- How do you ensure privacy of the data while providing broader access to the business?

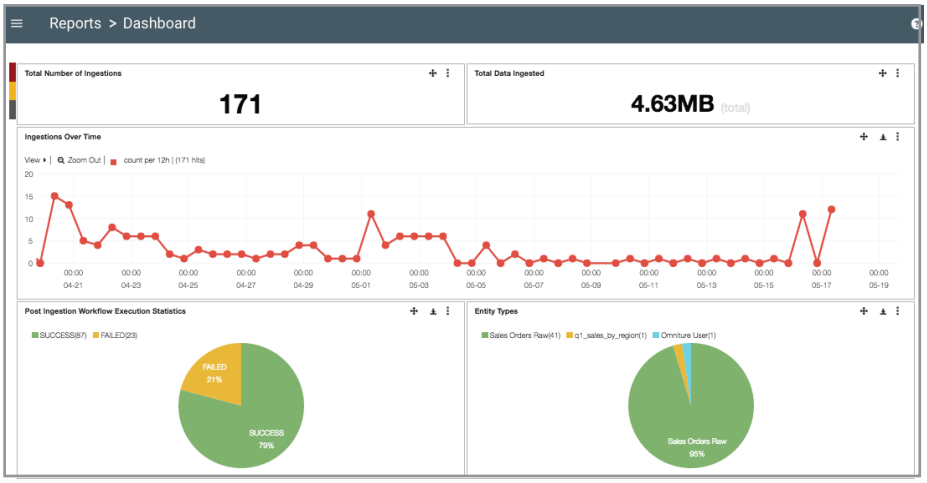
Let's look at this in more detail through the lens of the end-to-end data pipeline.

Step 1: Ingest

Ingest vast amounts of data from any source

Use a data lake management platform to define, track and log all steps of data ingestion in advance. You want this process to be repeatable and scalable and available for both file and stream ingestion. You should be able to configure the data lake management platform to automatically consume incoming files and

Ingest: Hydrate the lake with files and streams. Leverage a data lake management platform to automate and process new data



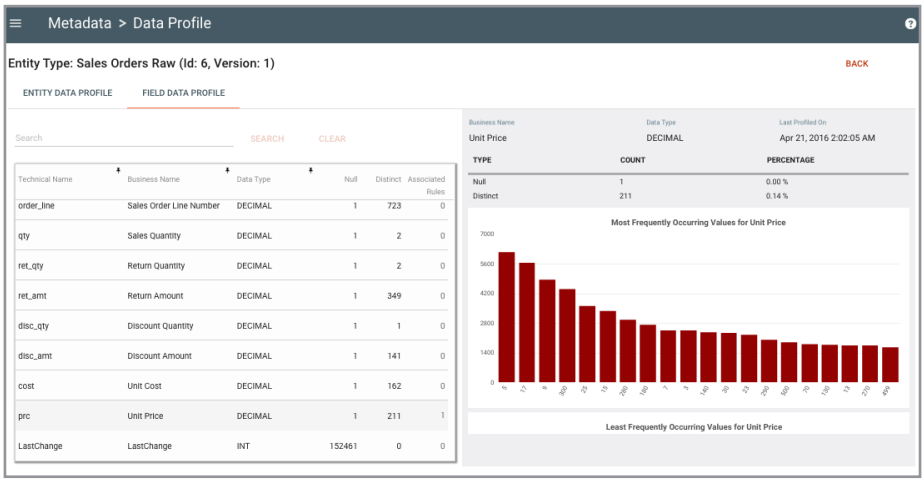
Example 1 - Ingest: The above screenshot demonstrates how a data lake management platform, such as Zaloni's Bedrock, can provide a dashboard to show the operations team onboarding of new data sets, managed so that IT knows where data comes from and where it lands.

Step 2: Organize

Know what is in your data lake

A data lake management platform is essential for organizing and managing data. It captures operational, technical and business metadata upon ingestion. You should be able to search, browse, and find the data you need for analytics, reducing your time to insight. Through file- and record-level watermarking, you should be able to see data lineage, where data moves and how it is used.

Organize: Automatically capture operational, technical and business metadata upon ingestion



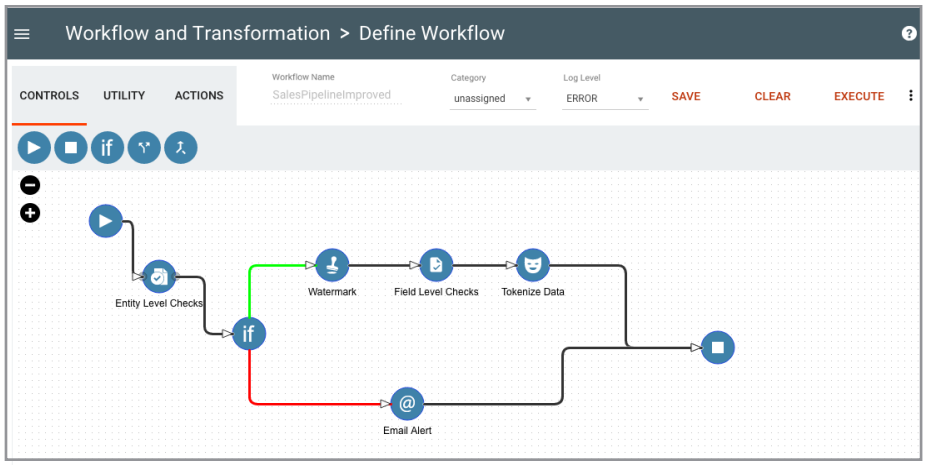
Example 2 - Organize: The above screenshot demonstrates how an integrated data lake management platform, like Zaloni's Bedrock, can show detailed summary views of the data in the data lake. These summaries show an operation team the types of data and the data range of the entities managed in the data lake.

Step 3: Enrich

Orchestrate and automate data preparation for actionable data

A data lake management platform simplifies data preparation by orchestrating workflows that deliver data security, quality and visibility. The platform should enable data tagging, masking and tokenization. It should allow you to convert data formats and orchestrate complex workflows to integrate updated or changed data. For example, our Spark-based transformation libraries provide flexible transformations at scale.

Enrich: Leverage a data lake management platform to orchestrate and automate data preparation



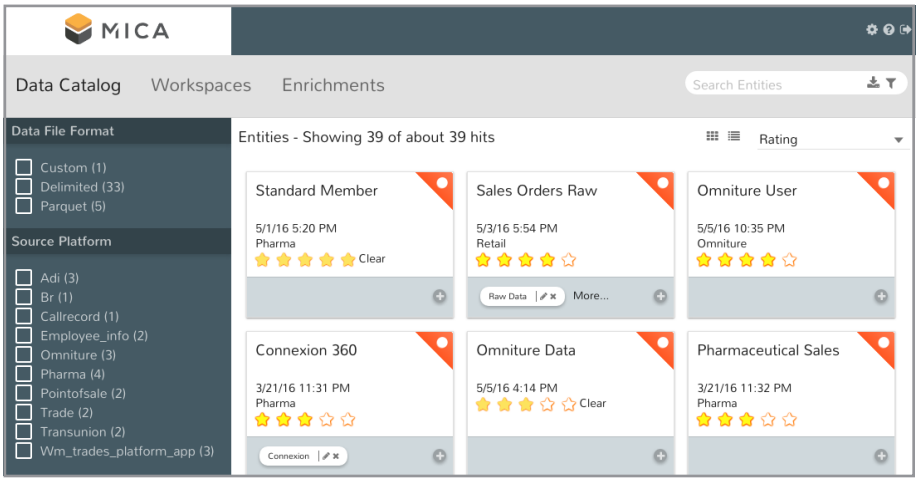
Example 3 - Workflows: The above screenshot demonstrates how an integrated data lake management platform, like Zaloni's Bedrock, can schedule conditional workflows on the entities managed by the data lake. Data quality checks can be run, and operations such as watermarking and tokenization can be performed and email alerts can be sent out on defined failure conditions.

Step 4: Engage

Democratize access to the data lake

The combination of a data lake management platform and a self-service data preparation tool can deliver cleansed, enriched and governed analytic-ready data sets to business end users. A data catalog is an essential tool for bringing information together in a single access point that lets you provision it for visualization tools like Qlik Sense and Tableau. A good self-service data preparation tool should be intuitive and provide a sandbox for less technical users to work with sample data from the dataset. Once the desired results are achieved, you should be able to put entire datasets into production and operationalize it to create repeatable, efficient processes for deploying datasets into production.

Engage: The combination of a big data management platform and a self-service data preparation tool can deliver cleansed, enriched and governed analytic-ready data sets to business end users



Example 4 - Engage: The above screenshot demonstrates how a data catalog and self-service data preparation tool, like Zaloni's Mica, can make the metadata readily available to business analysts and other stakeholders within the company. This data catalog tool stops the data lake from becoming a data swamp by showing what data is in the datalake and making the metadata easily searchable.

Most common use cases for managed data lakes today

Although the potential use cases are limitless, today managed data lakes are seeing success in these common use cases:

1. **EDW augmentation:** Offloading data from a traditional enterprise data warehouse (EDW) to Hadoop or the cloud brings storage cost savings and increases bandwidth in the EDW for business intelligence (BI) processes.
2. **Agile analytics:** A "fail fast" approach to data science where hypotheses, testing, iterating and improvements are in a constant cycle using real-time data, which can result in more and innovative insights that can add business value.
3. **Enterprise reporting:** The ability to do ad hoc reporting using an enterprise-wide data source is key to understanding the business in real-time and reducing risk.
4. **Data monetization:** More enterprises are leveraging their data to better understand current customers and also develop new products and services that resonate with consumers.
5. **Data science:** Some enterprises are working to support an enterprise-wide data science capability, particularly for predictive modeling and analytics using machine-learning and large datasets.

Common Managed Data Lake Use Cases:

1. EDW Augmentation
2. Agile Analytics
3. Enterprise Reporting
4. Data Monetization
5. Data Science

Managed data lake success: Case studies

Companies are making serious strides with big data implementations, addressing old problems in new ways and creating new opportunities to enhance their business, their customer loyalty and their competitive advantage. Below we outline 4 distinct use cases in 4 different industries:

1. Fraud detection in the insurance industry
2. EDW augmentation in market research
3. Hospital re-admission risk reduction in healthcare
4. Network analytics in telecommunications

1. Improving fraud detection and reduction

Workers' compensation fraud, including employee, employer and medical provider fraud, is estimated to cost states $1 billion-$3 billion each annually. One reason is that many fraudsters are able to manipulate billing faster than investigators can audit. To support a new real-time anti-fraud analytics solution for one of the country's largest provider of workers' compensation insurance, Zaloni built a data lake solution.

Before the data lake: The company's business analysts were using manual processes to build and run SQL queries from several systems, which could take hours or days to get results. The company wanted to unify its data silos and legacy data platforms to support a new anti-fraud analytics solution and enable non-

At a Glance:

- Industry: Insurance
- Company Description: Large US-based provider of workers' compensation
- Technical Use Case: Data Lake 360˚: Agile Analytics
- Business Use Case: Fraud and Payment Integrity
- Big Data Technologies: Zaloni Professional Services, Cloudera
- Deployment: On premises

Results: The new system improved data quality and the company's ability to detect patterns of potential fraud hidden in vast amounts of claims data, and enabled attorneys to quickly build cases with original claim documents. Data was used to prosecute more than $150 million in fraudulent cases. With queries now taking seconds versus hours or days, the company was able to analyze more than 10 million claims, 100 million bill line item details and other records.

2. Enjoying more cost-effective storage and faster data processing

A leading provider of market and shopper information, predictive analytics and business intelligence to 95% of the Fortune 500 consumer packaged goods (CPG) and retail companies needed a more cost-effective and efficient approach to process, analyze and manage data. Zaloni designed and built the backend solution architecture for an enterprise data warehouse (EDW) augmentation solution to provide a more cost-effective, flexible and expandable data processing and storage environment.

Before the data lake: Tasked with managing massive volumes of data (ingesting hundreds of gigabytes of data from external data providers every week), the company struggled to keep costs as low while providing clients with state-of-the-art analytic and intelligence tools.

With the data lake: Using Hadoop for the aggregate POS dataset and servicing the extractions that populate the company's custom, in-memory analytics farm enabled the company to offload more data, faster, and realize substantial savings in a very short period of time.

Results: The company realized $5.2 million annual savings and $4.4 million projected additional savings (upon completion of the fine-grained data-at-rest modification project). Additionally, the company saw a nearly 50% reduction in mainframe MIPS (millions of instructions per second/processing power and CPU consumption) and better throughput with Zaloni Bedrock managing the data lake.

At a Glance:

- Industry: Market Research
- Company Description: Top 10 Market Research Firm
- Technical Use Case: EDW Augmentation
- Big Data Technologies: Zaloni Bedrock, MapR, Zookeeper
- Deployment: On premises

3. Reducing re-admission risk

One of the US' leading healthcare companies wanted to more effectively use big data to deliver the best care more affordably. Specifically, it wanted to be able to better determine readmission risk for patient discharges and predictively classify readmissions as potentially preventable or not preventable. Zaloni helped them achieve this by developing a Hadoop-based managed data lake housing more than 60 million records of historical patient data.

Before the data lake: The company was unable to seamlessly access high volumes of both new and existing data from disparate sources and analyze it for patterns and trends.

With the data lake: The healthcare customer had a clear view of what data was in the data lake; could track its source, format and lineage; and enable users to more efficiently search, browse and find the data needed for analytics.

At a Glance:

- Industry: Healthcare
- Company Description: Healthcare Provider and Health Plan
- Technical Use Case: Data Lake 360°: Agile Analytics
- Business Use Case: Population Health: Re-admission Risk
- Big Data Technologies: Zaloni PS, Cloudera, MapReduce
- Deployment: On premises

Results: Zaloni's solution enabled the healthcare customer to reduce costs and improve outcomes by ultimately lowering readmission rates, due to significantly improved understanding of potentially preventable readmissions, the ability to develop better algorithms to more accurately predict readmissions, and a new ability to identify patients with the highest risk of readmission early in their initial hospitalization and proactively adjust treatment plans sooner to account for that risk.

4. Reporting compliance and improving customer experience

A large telecom company was required to archive large and growing volumes of wireless call details records (CDRs) to comply with government regulations. Although it was a significant cost to store all of this data, the company identified an opportunity to leverage the data to provide better customer service, grow its business and ensure it was getting the expected return on billions invested in capital expenses. To enable this goal, Zaloni architected and built a managed data lake to serve as the single repository for all traffic-related, inventory and provisioning data (CDRs, SNMP, server logs).

Before the data lake: The company's wireless network generated 4 terabytes of data per day from voice, data and SMS CDRs. This data was created by 11 different servers and switches with 8-10 different record layouts. As a result, there was no centralized repository to store all these CDRs. In addition, the upstream mediation system, which is responsible for merging CDR records into a single record for each session, was sending duplicate CDRs or not stitching the call records completely.

With the data lake: The company was better able to perform load balancing, government reporting, network analysis, and parallel processing that removed duplicate data and reconstructed call records from partial or lost records. With a managed data lake, the telecom company not only gained a compliance solution, but was able to more efficiently manage the network and continue to provide a great customer experience as the volume of subscriber usage continued to grow significantly.

Results: The company was able to double data ingestion, to 8 terabytes/day. Also, the company avoided costly fines and penalties by meeting the immediate need for government reporting requirements, and gained insights into network utilization to avoid network congestion in near real-time.

At a Glance:

- Industry: Telecommunications
- Company Description: South American telecommunications giant
- Technical Use Case: Data Lake 360°: Agile Analytics
- Business Use Case: NOC/SOC: Network Analytics
- Big Data Technologies: Zaloni Professional Services, Pivotal, MapReduce, Hive
- Deployment: On premises

Time to take the next step

It's a good idea to find the right tools and expertise that will help you achieve success in transitioning to a modern data architecture. Zaloni provides a Data Lake 360° solution that allows customers to leverage our experience with many data lake implementations. Customers also benefit from our data lake management platforms and tools, our handle on the quickly evolving big data ecosystem, and the specialized technical skills we bring to each project. Most importantly, we help ensure that enterprises truly derive value from their data lakes, helping to implement specific business use cases and making big data management and analytics more efficient and cost-effective.

BEDROCK

Data Lake Management Platform

Bedrock is a fully integrated data lake management platform that provides visibility, governance, and reliability. By simplifying and automating common data management tasks, customers can focus time and resources on building the insights and analytics that drive their business.

MICA

Self Service Data Preparation

Mica provides the on-ramp for self-service data discovery, curation, and governance of data in the data lake. Mica provides business users with an enterprise-wide data catalog through which to discover data sets, interact with them and derive real business insights.

Zaloni Professional Services

Your trusted partner for building production data lakes

Zaloni has more than 400+ staff years of big data experience working globally across the US, Latin America, Europe, Middle East and Asia. Zaloni Professional services offers expert big data consulting and training services, helping clients plan, prepare, implement and deploy data lake solutions.

Professional Services Include:

- Big Data Use Case Discovery and Definition
- Data Lake Assessment Services
- Solution Architecture Services
- Data Lake Build Services
- Data Lake Analytics Application Development
- Data Science Services

About Zaloni

Delivering on the Business of Big Data

Zaloni is a provider of enterprise data lake management solutions. Our software platforms, Bedrock and Mica, enable customers to gain competitive advantage through organized, actionable big data lakes. Serving the Fortune 500, Zaloni has helped its customers build production implementations at many of the world's leading companies.

To learn more about Data Lake 360°:

Call us: +1 919.323.4050 | **E-mail:** info@zaloni.com | **Visit:** www.zaloni.com/data-lake-360

Find Us on Social Media:

Twitter handle @zaloni