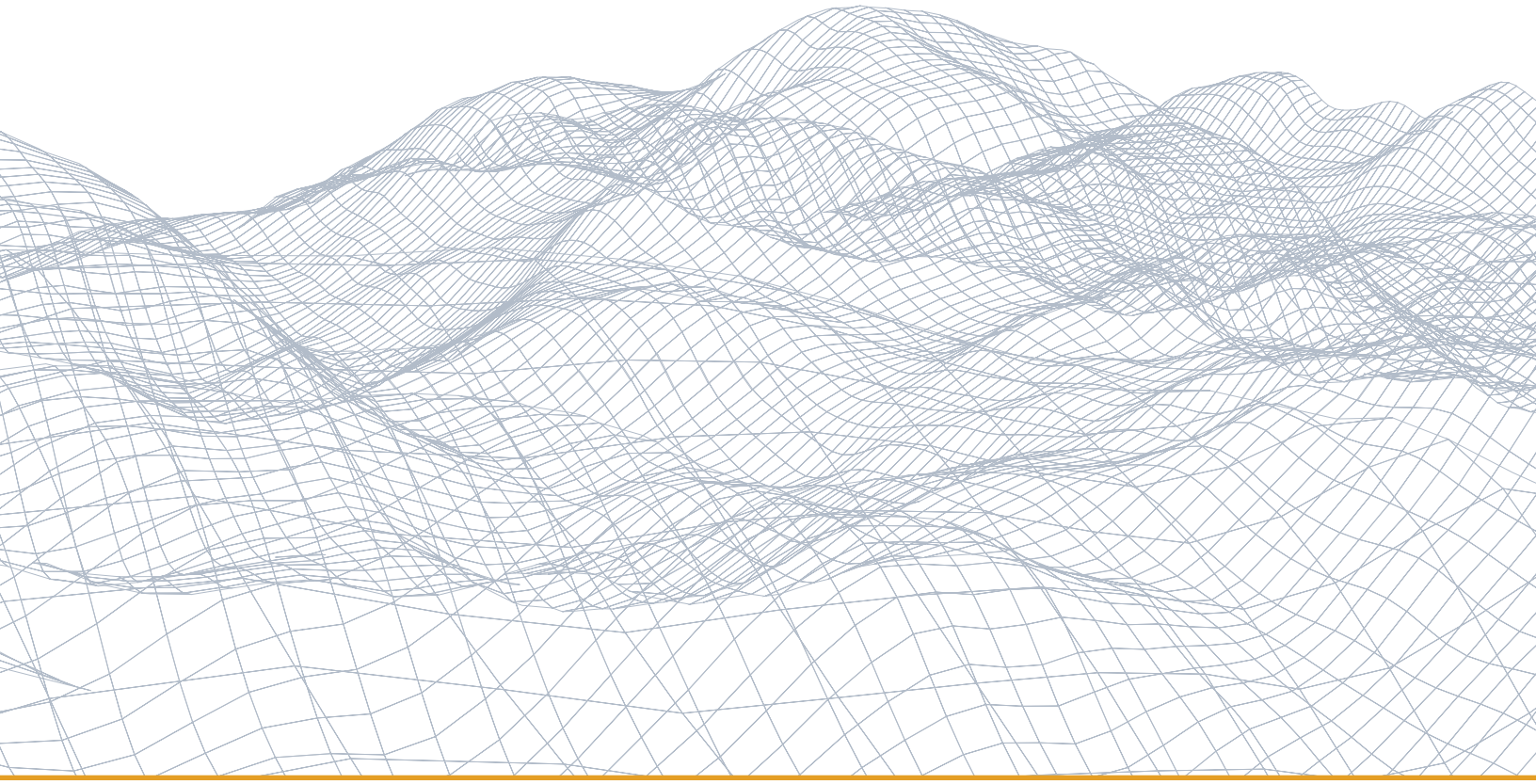
Why Your Data Warehouse Needs a Data Lake and How to Make Them Work Together

Strategies and best practices for implementing a successful DW augmentation



ZALONI
THE DATA LAKE COMPANY

www.zaloni.com

Introduction

Managed data lakes are the wave of the not-so-distant future, destined to become central components of the modern data architecture. They provide the flexibility, scalability and agility required by enterprises to manage the volume, types, and real-time availability of data that is generated today. They also can democratize data access for more users for nearly any purpose.

But for many businesses, such a significant shift in overall data architecture (and culture) may not be a realistic option quite yet. However, as data – and the speed and volume at which it is collected and analyzed – becomes more critical to business outcomes and operations, it's important to start thinking about how to begin to position your company now so that it's not left flat-footed in the big data race.

Complementing your existing data warehouse (DW) with a data lake can be a smart first step for many companies.

Complementing your existing enterprise data warehouse (DW) with a data lake can be a smart first step for many companies. It gives you more flexibility and speed in terms of data processing and capturing unstructured, semi-structured and streaming data, and frees up bandwidth in your data warehouse for business intelligence analytics. It's also a use case that typically produces a guaranteed return on investment.

What we mean by "data lake" and DW "augmentation"

When we say "data lake," we're referring to a centralized repository, typically in Hadoop, for large volumes of raw data of any type from multiple sources. It's an environment where data can be transformed, cleaned and manipulated by data scientists and business users. A "managed" data lake is one that uses a data lake management platform to manage ingestion, apply metadata and enable data governance so that you know what's in the lake and can use the data with confidence.

"Augmentation" means enhancing what you already have, not starting from scratch. With a data warehouse augmentation, you keep your data warehouse and existing BI tools, but add a complementary, integrated data lake. For example, you could use a complementary data lake to prepare datasets and then feed them back into a traditional data warehouse for business intelligence analysis, or to other visualization tools for data science, data discovery, analytics, predictive modeling and reporting.

Why DW augmentation?

We find that companies typically consider a DW augmentation project for two scenarios:

1. **Blue sky** – You want to be able to do new things, beyond the capabilities of the data warehouse. This could include supporting specific business use cases for more advanced big data analytics or data science to find new insights or generate revenue; for example, with new products and services or through improved, more personalized customer experience.
2. **Cut costs** – You want to continue doing what you're already doing with your data warehouse, but do it cheaper using commodity hardware.

In either case, adding a data lake gives you major benefits in terms of speed and flexibility. Data moves through a data lake much faster than a data warehouse, reducing latency and providing quicker time to insight. Data lakes also can support streaming so that data is continuous and not relegated to periodic updates, as with batch-based data warehouses. And, as you know, the structure of a data warehouse is difficult to change as the needs of your business change. A data lake is much more flexible when it comes to keeping up with dynamic business rules and data needs.

Why companies typically consider a DW augmentation project:

- Blue sky - to be able to do new things beyond the data warehouse
- Cut costs - reducing costs by leveraging commodity hardware

Approaches: A framework for a data warehouse agumentation

We're not going to sugar coat it: data lake integrations with existing architectures and parts of your business can be pretty demanding, depending on your goals. However, by having a comprehensive understanding of key challenges and putting a strategy in place (and maybe working with a few experts), you are sure to have success.

Let's talk strategy

With an DW augmentation, you're combining traditional technologies with new technologies. The strategy is to create an environment in which you use the technology that is the best fit for the job, versus trying to retrofit your existing software that either can't do the job as well or requires your team to spend valuable time and resources to try and make it work. For example, software you've used to import data to your data warehouse could be used when importing data into the data lake, but extraction, transformation and loading (ETL) will happen outside the Hadoop cluster. Instead, moving ETL to the data lake can be faster and cheaper -- allowing your team to spend less time on data preparation and more time on more valuable activities like analytics.

It's important to note that a data lake is a new platform, not just another database. It's a different architecture with different hardware, software and operations. The most striking difference is that unlike having to process data to fit a data warehouse's predefined schema, the data lake ingests raw data in its native format. That means that unlike the data warehouse, data can be ingested into the data lake regardless of its quality or completeness. You need to use a data management platform in order to apply metadata so that you can understand the quality of your data.

Below is a comparison chart that highlights several other key differences.

Attribute	DW	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 processing, plus an improved capability over DWs to support big data inquires 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
Benefits	• Transform once, use many • Clean, safe, secure data • Provides a single enterprise-wide view of data from multiple sources • Easy to consume data • High concurrency • Fast response times • Mature governance • Operational integration	• Transforms the economics of storing large amounts of data • Provides a single enterprise-wide view of data • Scales to execute on tens of thousands of servers • Easy to consume data • Allows use of any tool • Enables analysis to begin as soon as data arrives • Fast response times • Allows usage of structured and unstructured content form a single source • Supports agile modeling by allowing users to change models, applications and queries • Mature governance • Analytics and big data analytics

The most striking difference between a data lake and a data warehouse is the ability to ingest data in its native format providing flexibility and allowing a far greater and timelier stream of data for analysis.

A data lake requires a different approach to security and compliance, as the security features built into the data warehouse platform don't exist in Hadoop. Furthermore, the security tools you may already be using for your data warehouse can't scale to keep up with big data, or just won't work at all. While Hadoop's security is evolving and certainly has improved over the last few years, third-party security solutions are required to meet compliance regulations in areas such as encryption at rest, user authentication, data access history tracking and authorization. Just be aware that configuring all the mechanisms and settings can be a complex task.

Data governance challenges

Governance is not automatically applied in a data lake. You have to apply rules for how data is managed and tracked (e.g., categorized as raw, refined, trusted, watermarked, etc.), protected for privacy (e.g., masking, tokenization) and accessed (e.g., permissions for who can use what data) throughout the data pipeline. Also, governance in the data lake is different than for the data warehouse. In the data lake, governance rules can be flexible, based on the type of data that is being ingested. For example, some data could be certified as accurate and of high quality, while other data might require less accuracy to be useful and therefore require different governance rules and controls.

The importance of automation

As your data lake grows over time, leveraging automation is the only way you can successfully operate at the scale of big data. In particular, automation is important for functions such as data ingestion, metadata management and data lifecycle management. Automation reduces errors, increases speed – and makes management and tracking of hundreds or thousands of data sources feasible.

Automation reduces errors, increases speed and makes management and tracking of hundreds or thousands of data sources feasible.

Metadata is the key to the castle

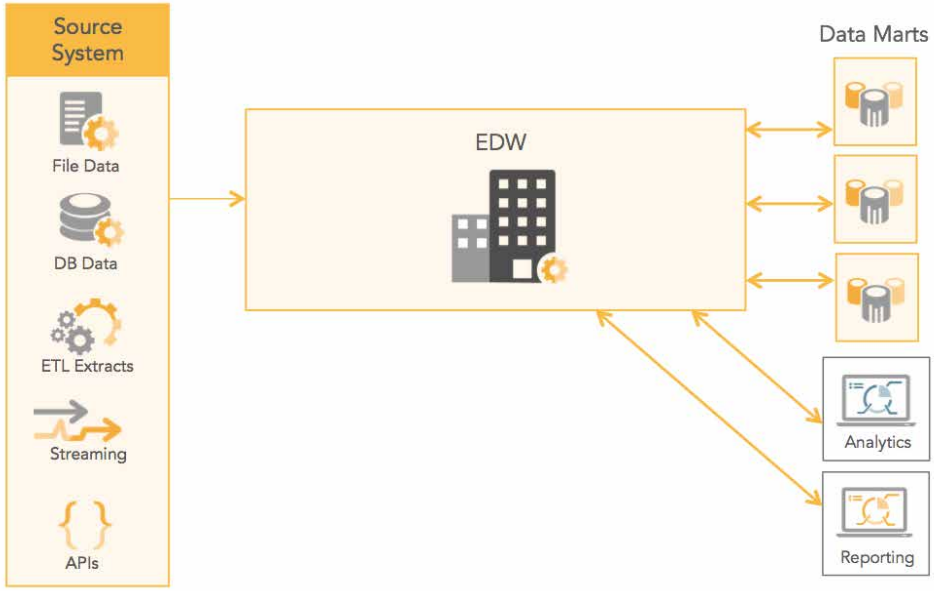
Metadata is what allows you to derive value from your data lake. Without it, you can't find the data you want, use it with confidence, or apply data governance rules. It is absolutely critical to automate the application of metadata to your data upon ingestion into the data lake. We recommend applying three types of metadata to have the most complete picture of your data.

Type of Metadata	Description	Example
Technical	Captures the form and structure of each data set	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, lineage
Business	Captures what it all means to the user	Business names, descriptions, tags, quality and masking rules

Reference Architecture

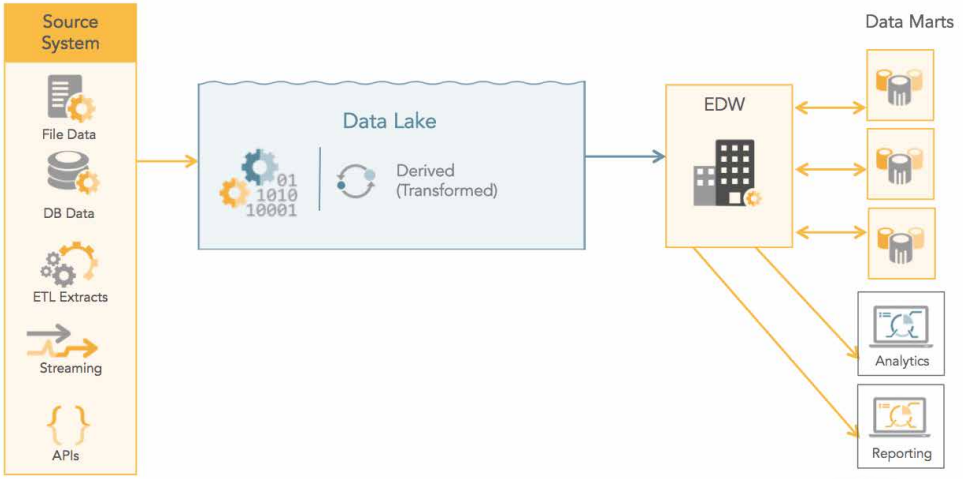
What could a data warehouse augmentation look like in your environment? Let's review some architecture diagrams.

Reference Diagram #1: Traditional data warehouse reference architecture



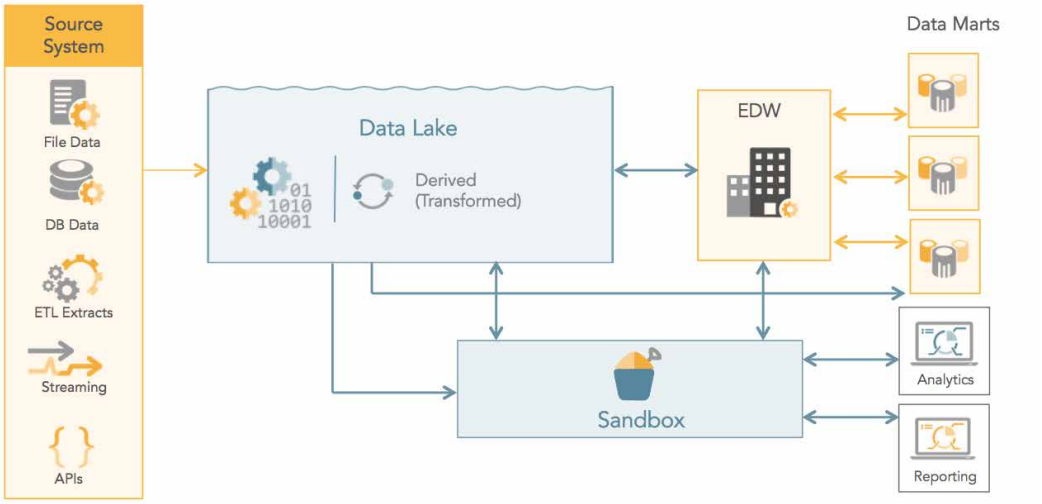
The differences between a traditional data warehouse architecture and data lakes are significant. An DW is fed data from a broad variety of enterprise applications. Naturally, each application's data has its own schema, requiring the data to be transformed to conform to the DW's own predetermined schema. Designed to collect only data that is controlled for quality and conforming to an enterprise data model, the DW is capable of answering only a limited number of questions. Further, storing and processing all data in the data warehouse is cost prohibitive.

Reference Diagram #2: Data warehouse that has been augmented with a data lake



Typically an organization will augment a data warehouse with a data lake in order to enjoy a reduction in storage costs. The data lake is fed information in its native form and little or no processing is performed for adapting the structure to an enterprise schema. The data can be stored on commodity hardware, rather than expensive proprietary hardware. Required data can be pulled from the lake to leverage in the data warehouse. While this model provides significant cost-savings it does not take advantage of the strategic business improvements that a data lake can provide.

Reference Diagram #3: Data Warehouse Augmentation and Offload Diagram



The biggest advantage of data lakes is flexibility. By allowing the data to remain in its native format, a far greater stream of data is available for analysis. When an organization enlists the data lake to offload expensive data processing, in addition to storage, the entire business can benefit from more timely access to more data.

Business intelligence/visualization tools integration

When you have existing business intelligence (BI) tools that require low latency, the data lake ecosystem can quickly become complex and harder to manage. Incorporating streaming data for analytics can require integration of up to 10 components, such as Apache Kafka, REST APIs, Java database connections and others to manage ingestion, enrichment, data quality, etc. -- and require that your team learn how to use each one.

Leveraging the cloud

Hadoop in the cloud allows for added flexibility. With elasticity and unlimited scalability, cloud services like AWS EMR or Microsoft Azure HD Insight allow you to spin up and scale Hadoop clusters on demand to support capabilities beyond what your data warehouse can offer. Also, storage and compute services are decoupled, so you can pay for storage at a lower rate than for computing services. Also, if you use a comprehensive data lake management platform, you can take advantage of transient clusters, which are compute clusters that automatically shut down and stop billing when processing is finished (versus persistent compute).

Although metadata is automatically deleted by the cloud provider when a transient cluster is shut down, a data lake management platform can maintain the metadata outside the cluster, making it available after the cluster is terminated.

Lastly, know that it doesn't have to be all or nothing when it comes to the cloud. Many companies take a hybrid approach. A data lake management platform can help you manage a data lake that spans on-premises and cloud.

Your next steps: Taking Stock

Before tackling an EDW augmentation project, it's a good idea to review your current resources and understand what you'll need. If you decide to deploy a proof-of-concept data lake yourself, we recommend you at least have an expert do an architecture review before you go into production.

- ☐ Does my team have the right skill set and knowledge, including Java, Hadoop, and other NoSQL data sources?
- ☐ Is my team familiar with cloud technologies?
- ☐ Will my team be able to stay up-to-date with Hadoop's constantly changing ecosystem of tools?
- ☐ Is my team prepared to go through what can be an involved process of tool selection?
- ☐ Do we have business justification?

In practice: DW augmentation case studies

We have worked with a number of enterprises that have successfully implemented an EDW augmentation project leveraging Bedrock, Zaloni's data lake management platform – often as part of a larger initiative.

Media and Entertainment: Enterprise-wide data catalog

A major American multinational media and entertainment company wanted to give a range of data users self-service access to its enterprise-wide data. The intent was to enable faster time to insight into some of its metrics such as volume, quality, and business KPIs. Ultimately, the company was looking for a solution to extend its Teradata platform to include a data lake in order to help it shift to becoming a more data-driven organization across the enterprise.

DW Augmentation solution: We helped the company ingest data from all new data sources directly into the data lake and integrated all data sources for a single view with an enterprise-wide data catalog. Zaloni's self-service data preparation platform, Mica, made it easy for end users to search for datasets and refine, transform, enrich and extract data.

DW Augmentation Case Studies:

- Enterprise-wide data catalog
- DW offload
- Advanced analytics

Mica leveraged Bedrock for data organization, metadata management and operationalization of the execution of logic on the Hadoop cluster.

Results: The customer successfully migrated more than 20 of its critical datasets into the data lake. Further, the Mica self-service platform empowered the analytics and business intelligence teams, formerly dependent on IT or support teams, to create and prepare the data themselves.

Resort and Casino: Data warehouse offload

California's largest resort and casino manages a comprehensive customer loyalty program. The company wanted to be able to use existing and future data to optimize and grow its loyalty program by leveraging data science and advanced analytics.

DW Augmentation solution: The resort and casino's automated and configurable ingestion framework ingested data into the data lake from approximately 5,000 tables from Oracle and the SQLServer within weeks – compared to the months it would have taken to do it manually. Bedrock tracked lineage and metrics of every ingestion, allowing for a wider range of analytics, faster. Data could be accessed for self-service analytics in near real-time, or stored in raw format to be used if it became important in the future.

Results: The resort and casino was able to prove the potential value of specific use cases, convincing its board of directors to move forward into production, with the end goal of optimizing the customer loyalty program and improving the customer experience. In addition, the solution provided a framework for the future onboarding of other sources without needing additional development.

Financial Information and Analytics: Advanced analytics

A leading provider of independent ratings, benchmarks, analytics and data was looking for a big data solution to enable more advanced, real-time real estate industry analytics.

DW Augmentation solution: The solution used WebHDFS for ingestion, Apache Hive for processing and Apache Spark and for data enrichment. Data was able to be pushed back to the data warehouse allowing the company to run traditional reporting from their SQL server.

Results: Zaloni migrated a dataset 150+ million rows and 125 columns wide, which was growing 10% each month, from the company's data warehouse to Hadoop.

Building your big data future

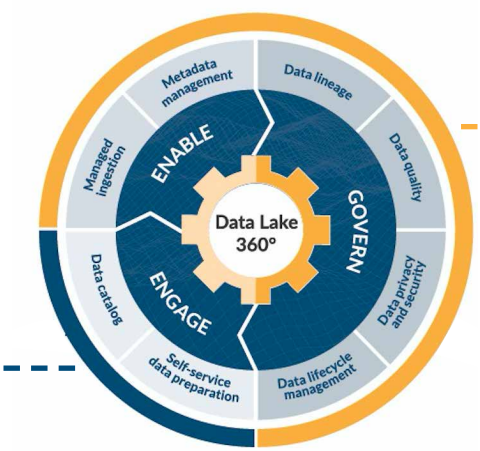
Ideally the data lake enables companies to do much more than optimize their data warehouse, although an EDW augmentation is a good first step. Modern data architectures and tools that can accommodate and analyze data are changing the ways enterprises can derive value from data. It is the inherent and 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.

Zaloni works closely with enterprises to design the architecture for data warehouse augmentations and implements Bedrock – the industry's only fully integrated Hadoop data lake management platform – to not only accelerate deployment, but also significantly improve visibility into the data.

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.



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.

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:

Call us: +1 919.323.4050 | E-mail: info@zaloni.com | Visit: www.zaloni.com

Find Us on Social Media:

Twitter handle @zaloni