Eckerson Group

The Definitive Guide to Data Warehouse Modernization

Data Management Patterns for Next-Generation Analytics and AI

By Dave Wells

July 2019

Research Sponsored by

Informatica

About the Author

Dave Wells is an advisory consultant, educator, and research analyst dedicated to building meaningful connections throughout the path from data to business impact. As an educator he has written dozens of courses and taught hundreds of classes about data warehousing, data modeling, data architecture, and business intelligence for professional organizations such as Dataversity, eLearningCurve, and TDWI.

Dave works at the intersection of information and business, driving value through analytics, business intelligence, and innovation. More than 40 years of information management experience combined with over 10 years of business management create a unique perspective about the connections among business, information, data, and technology. Knowledge sharing and skills building are Dave's passions, carried out through consulting, speaking, teaching, and writing.

About Eckerson Group

Eckerson Group helps organizations get more value from data and analytics. Our experts each have more than 25+ years of experience in the field. Data and analytics is all we do, and we're good at it! Our goal is to provide organizations with a cocoon of support on their data journeys. We do this through online content (thought leadership), expert onsite assistance (full-service consulting), and 30+ courses on data and analytics topics (educational workshops).

Get more value from your data. Put an expert on your side. Learn what Eckerson Group can do for you!

Full-Service Consulting
Educational Workshops
Thought Leadership

Get More Value From Your Data

The research for this report is made possible by **Informatica**.

Table of Contents

Executive Summary

Modern data management practices raise questions about the role of data warehousing. Though some have declared the data warehouse dead, many organizations operate at least one data warehouse (most have two to five) and expect to do so for the foreseeable future. Data warehousing continues to be an important part of data management, but we need to modernize aging data warehouses to fit gracefully into modern data management practices and to deliver sustained value into the foreseeable future. Legacy data warehouses must evolve both architecturally and technologically to fit into modern analytics ecosystems. This report provides guidance to sustain the value of data warehouses through architectural restructuring, cloud migration, and integration into a comprehensive and cohesive data management strategy.

Why Data Warehouse Modernization?

Despite declarations to the contrary, data warehousing is not obsolete. Recent polling shows that more than 60% of companies are operating between 2 and 5 data warehouses today. Fewer than 10% have only one data warehouse or no data warehouse at all. Nearly one-third of poll respondents work in an organization with 6 or more data warehouses. Although the vision from the past generation of BI and data warehousing—one data warehouse that serves as a single version of the truth—has not been realized, it is clear that data warehousing continues to provide value to these organizations.

Data warehousing is not dead, but it is struggling. It is alive, but perhaps not entirely well. Big data, data lakes, NoSQL, data science, self-service analytics, and demand for speed and agility all challenge legacy data warehousing. Traditional data warehousing—predicated on 1990s data management practices—simply can't keep up with the demands of rapidly growing data volumes, processing workloads, and data analysis use cases. Data warehousing must evolve and adapt to fit the realities of modern data management and to overcome the challenges of scalability, elasticity, data variety, data latency, adaptability, data silos, and data science compatibility.

Companies continue to operate existing data warehouses because they are needed. Business processes and information workers depend on warehouse data and information on a daily basis. Many people—perhaps the majority—continue to need well-integrated, systematically cleansed, easy-to-access relational data that includes a large body of time-variant history.

They want to meet routine information needs with data that is prepared and published with those needs in mind. The future of data warehousing depends on data warehouse modernization, including architectural rethinking and purposeful use of cloud technologies.

The Evolution of Data Warehousing

Data warehouse modernization (DWM) is a natural next step in the evolution of data management for modern analytics, artificial intelligence (AI) and machine learning (ML) projects. Warehousing originated to meet difficulties of non-integrated operational systems and the resulting data disparity. Data management architecture was linear, and the typical use cases were reporting and business intelligence. (See Figure 1.)

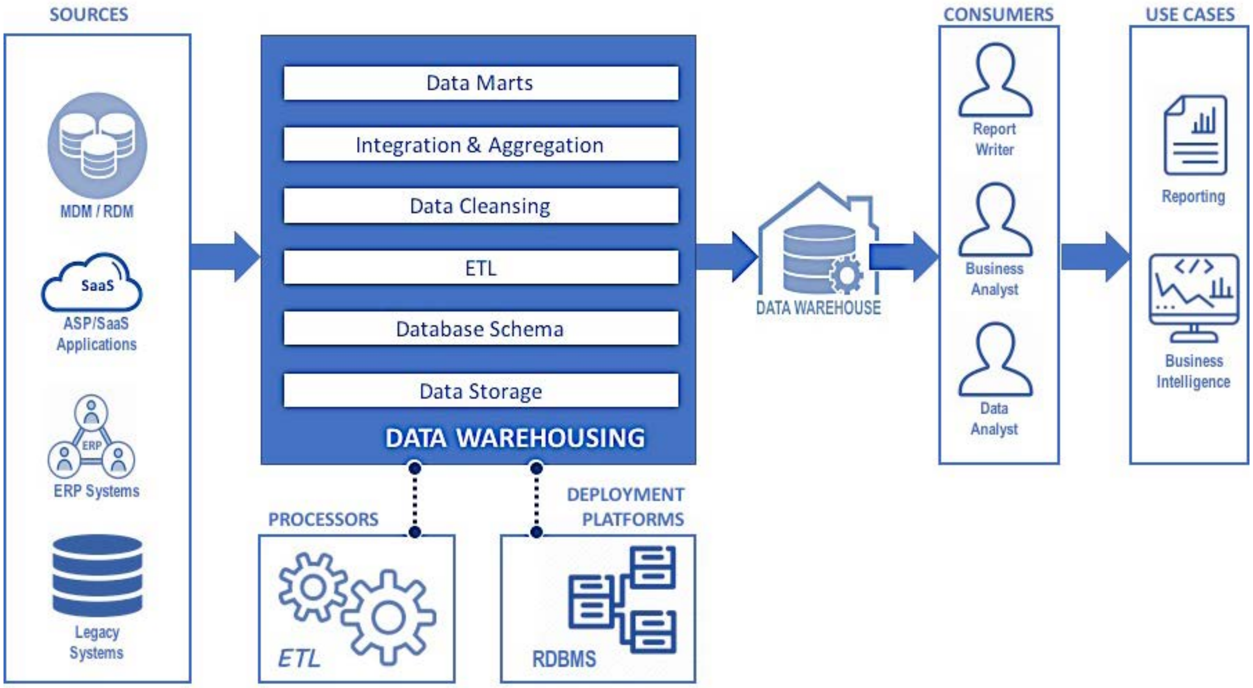


Figure 1. Legacy Data Warehousing Architecture.

This architecture served throughout the period when data sources were primarily internal structured data, and relational databases satisfied data storage and management needs. It readily adapted to support data marts, multidimensional data, and OLAP analysis.

But slowly the strength and stability of data warehouses eroded. Mergers, acquisitions, and other changes resulted in companies having multiple data warehouses—the next generation of data silos.

Then the age of big data arrived and disrupted long-standing data management practices. Legacy data warehouses are ill equipped to handle unstructured data complexities, process massive data volumes, adopt NoSQL databases, leverage the processing power of Hadoop, and harness the scalability and elasticity of cloud technologies.

Enter the data lake! Data lakes quickly became the next-generation concept for data management—optimized for big data, embracing NoSQL, powered by cloud technologies, and harnessing the power of open-source technologies such as Apache Spark, Kafka, and Hadoop. Data management utopia? Well … not quite. Data lakes didn't replace data warehouses. Together with multiple data warehouses we now have a new generation of data silos. (See Figure 2.)

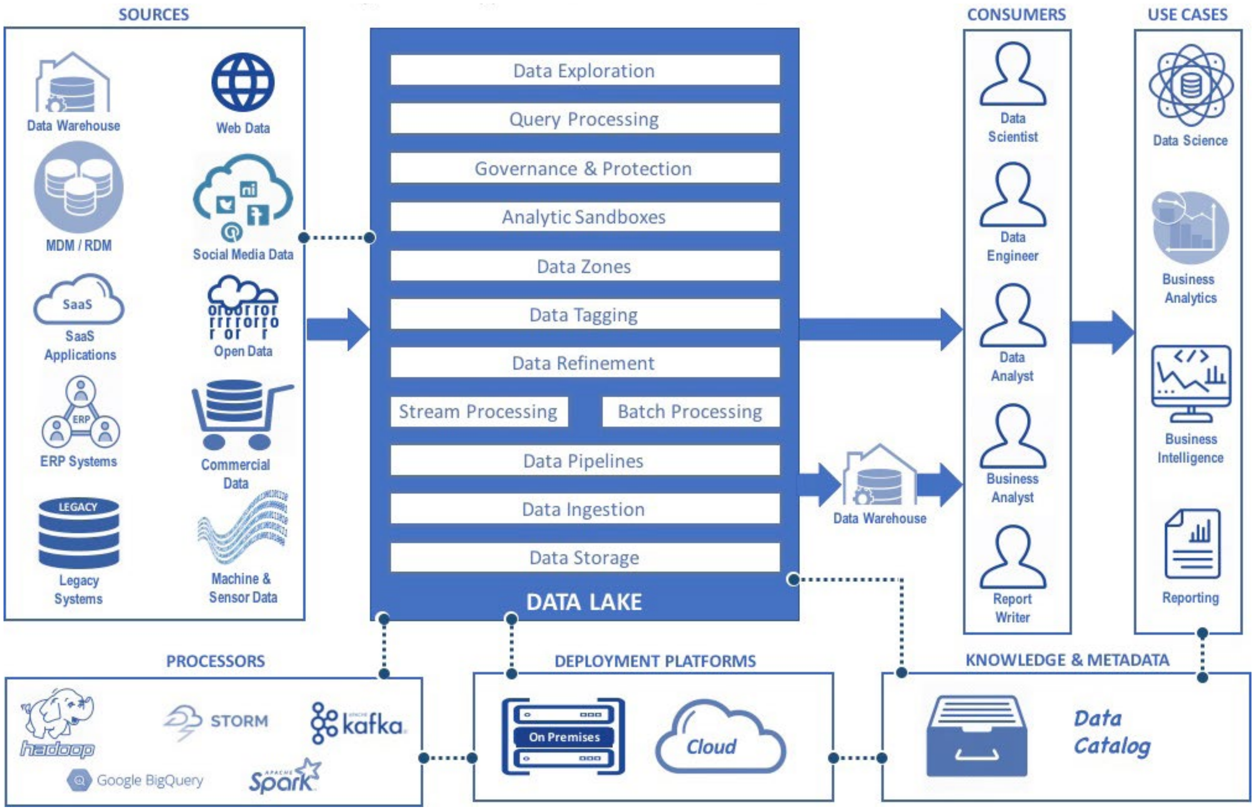


Figure 2. Typical Data Lake Architecture

Organizations with data lakes also experienced new challenges in adopting open source technologies. The multitude of these technologies, configuration complexities, and continuous change make software and infrastructure management difficult. These difficulties are compounded by lack of a skilled workforce outside of a few geographic locations such as Silicon Valley. Open source talent is scarce and costly, continuous retraining is essential, and employee retention is problematic in a highly competitive labor market.

Ten Must-Haves of Data Warehouse Modernization

Today's data lakes are not the end point of data management evolution. They do not render data warehousing obsolete, and they are but a step along the path toward a future of enterprise data hubs. Data warehouse modernization (DWM) is a natural next step in the evolution of data management. It is necessary to ensure sustained value of warehouse data and continued return on the investments already made in data warehousing. We need to rethink data warehousing architecture and data warehouse deployments now! The gap between legacy data warehousing and modern data management practices grows wider with each technology innovation.

Consider DWM as a key component of data strategy. It has an important role in shaping your data management future. Modernize now for a data warehousing future in which you can:

1. **Get full advantage from cloud technologies.**

    o Scalability—Horizontal scaling or scale-out adapts to changing workloads quickly.

    o Elasticity—The ability to expand and cut capacity as the workload fluctuates is especially important in data warehousing where data volumes, processing workload, and number of concurrent users can experience extreme peaks and valleys.

    o Managed infrastructure—Shifting the burden of data center management to the services provider and eliminating management workload for floor space, rack space, power, heating and cooling, and hardware and software management.

    o Cost savings—Reducing cost of operating an on-premises data center and shifting much of data management cost from capital expenditure to operating expense.

    o Processing speed—Cloud computing delivers significantly faster processing. Much of the gain is a direct result of scale-out architecture—the ability to add processing capacity horizontally and to expand and contract (elasticity) as needed.

✓ Scalability
✓ Elasticity
✓ Managed infrastructure
✓ Cost savings
✓ Processing speed
✓ Deployment speed
✓ Disaster recovery
✓ Security and governance

- Deployment speed—Data warehouse enhancements and modifications seem endless, but projects are frequently delayed for infrastructure upgrades to expand data capacity, increase processing capacity, or support additional development and test environments. Cloud elasticity overcomes these challenges to eliminate project delays and accelerate deployment.
- Disaster recovery—Business critical data warehouses are often overlooked in disaster recovery planning, in part because the complexities of warehousing make disaster recovery planning especially difficult. Virtualization in a cloud environment offers opportunity for a simpler approach.
- Security and governance—Of the many dimensions to data security, some are the full responsibility of cloud service providers. Server security, for example, becomes a provider responsibility when migrating a data warehouse to the cloud. Other security and governance aspects become shared responsibilities where understanding provider features and capabilities and describing responsibilities through SLAs is important.

2. **Support cloud-hybrid and multi-cloud environments.** As the deployment landscape expands, seamless interoperability across multiple technology environments becomes critical. The complex deployment landscape shown in Figure 3 illustrates the realities of typical deployments today. This landscape

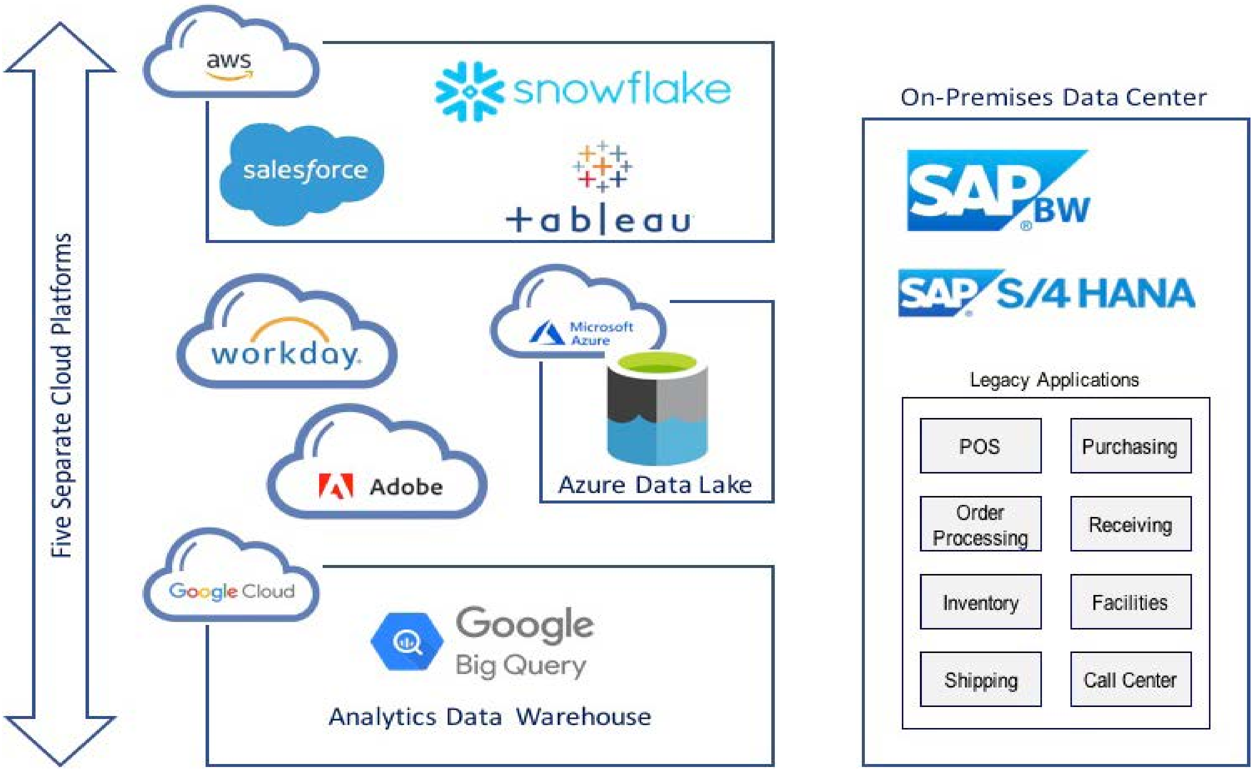


Figure 3. Complex Deployment Landscape

has four separate cloud environments as well as several systems operated in an on-premises data center. A modern data ecosystem must support cloud to on-premises interoperability such as connecting the Snowflake data warehouse with the SAP business warehouse or connecting legacy applications with the analytics data warehouse implemented in Google Cloud. The ecosystem must also support cloud to cloud interoperability such as connecting Workday applications with the Azure data lake or working simultaneously with the Snowflake data warehouse and the analytics data warehouse. Ultimately multiple cloud environments along with on-premises systems will be standard. Interoperability is critical as the systems must all work together without isolating any of the data that they store and manage.

3. **Support all types of data—structured, semi-structured, and unstructured**. The once simple world of structured data stored in relational tables has given way to the big data phenomenon. Modern data management continues to work with structured data such as customer records and sales transactions that are rigorously organized as rows and columns. Structured data in the Hadoop ecosystem sometimes shifts from relational tables to cloud-optimized and Hadoop-friendly formats such as Avro and Parquet. Avro is a row-based storage format and Parquet is a column-based format, both optimized for Hadoop. Semi-structured data is less rigorously organized and is typically stored with file formats that use semantic tagging such as XML and JSON. Machine generated data from sensors, mobile devices, and mobile apps is often collected and stored as semi-structured data. Semi-structured data formats are also a common way to share data through electronic data interchange (EDI) services. Unstructured data is at the opposite extreme, lacking the organization of structured data and the semantic context of semi-structured data. Unstructured data is often textual but also includes images, photos, and videos. Freeform text customer comments accompanying a warranty service request and photos associated with insurance claims are examples of unstructured data. Legacy data warehouses are bound by the relational constraints of structured data. Modern data warehouses must support all types of data.

4. **Support all data latencies—batch, real time, and streaming.** Legacy data warehouses are generally populated through batch extract-transform-load (ETL) processing, which is inherently high-latency. Daily loads, for example, create a data warehouse where the most current data is already one day old. Today's real-time business processes often demand real-time data. A modern data warehouse must support data at all speeds, continuing to use batch processing when appropriate, using changed data capture (CDC) techniques to acquire data in real time, and parsing data streams to capture only the events of interest.

Only then can the data warehouse support a wide range of common use cases including reporting for time-series analysis and trends, dashboards for real-time monitoring, and real-time alerts of business events and conditions discovered through data. (See figure 4.)

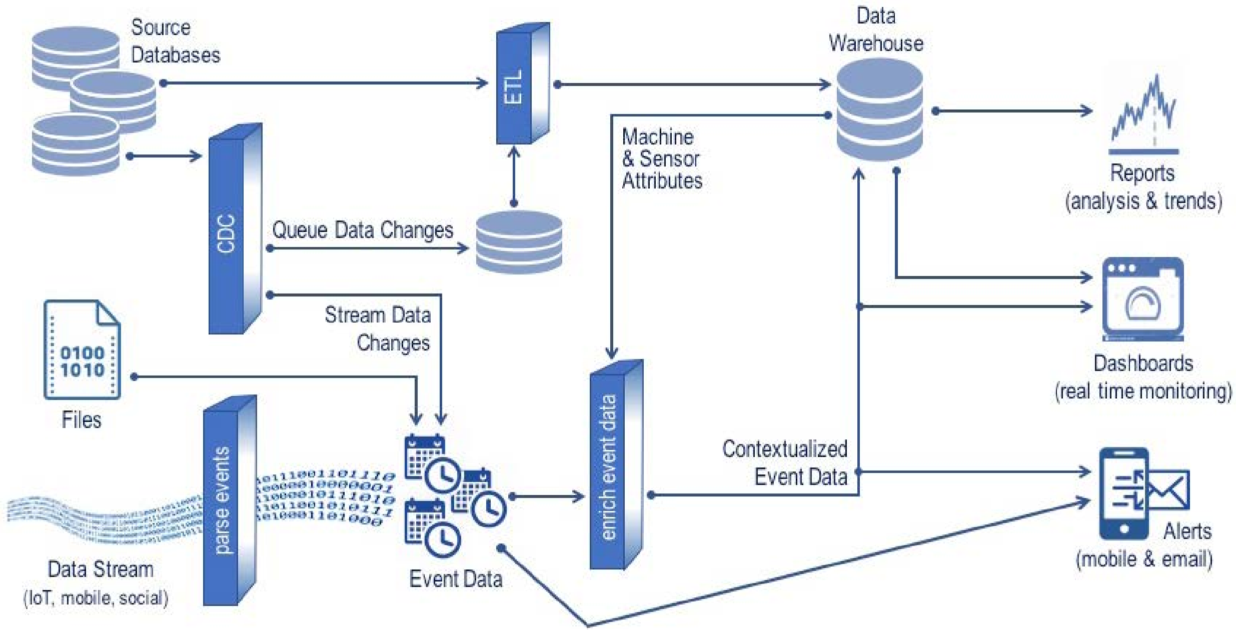


Figure 4. Support for All Data Latencies

5. **Support all data users—data scientists, data analysts, data engineers, and report writers.** Individuals in different roles have differing data needs. Data scientists often prefer raw data at atomic level of detail and without cleansing or other transformations applied. Data analysts—especially line-of-business analysts using self-service tools—get value from data that is integrated and cleansed, as it requires less data preparation work from them. Report writers prefer to work with data that is integrated, cleansed, dimensioned, and aggregated. Data engineers work with data in all of these forms. The modern data warehouse ideally supports all users with data ranging from raw to highly transformed, and with lineage and traceability throughout. (See figure 5.)

6. **Support collaboration among all data users.** Data-driven and collaboration go hand-in-hand as key traits of modern business culture. People working with data must work collaboratively to share knowledge, analysis, and data; never in isolation. (See figure 5.) Data scientists can create models that others build upon. Data engineers can create data preparation processes that can be reused. Data analysts can share their analysis for others to discover and use or adapt, eliminating the time and cost of redundant analysis. Every data user can

share knowledge about data and their experiences when working with specific datasets. Collaboration and sharing improves efficiency, improves quality of analysis and reporting, and elevates data literacy across the organization. Strong connections and a high level of interoperability between the data warehouse and the data catalog are key elements of collaboration.

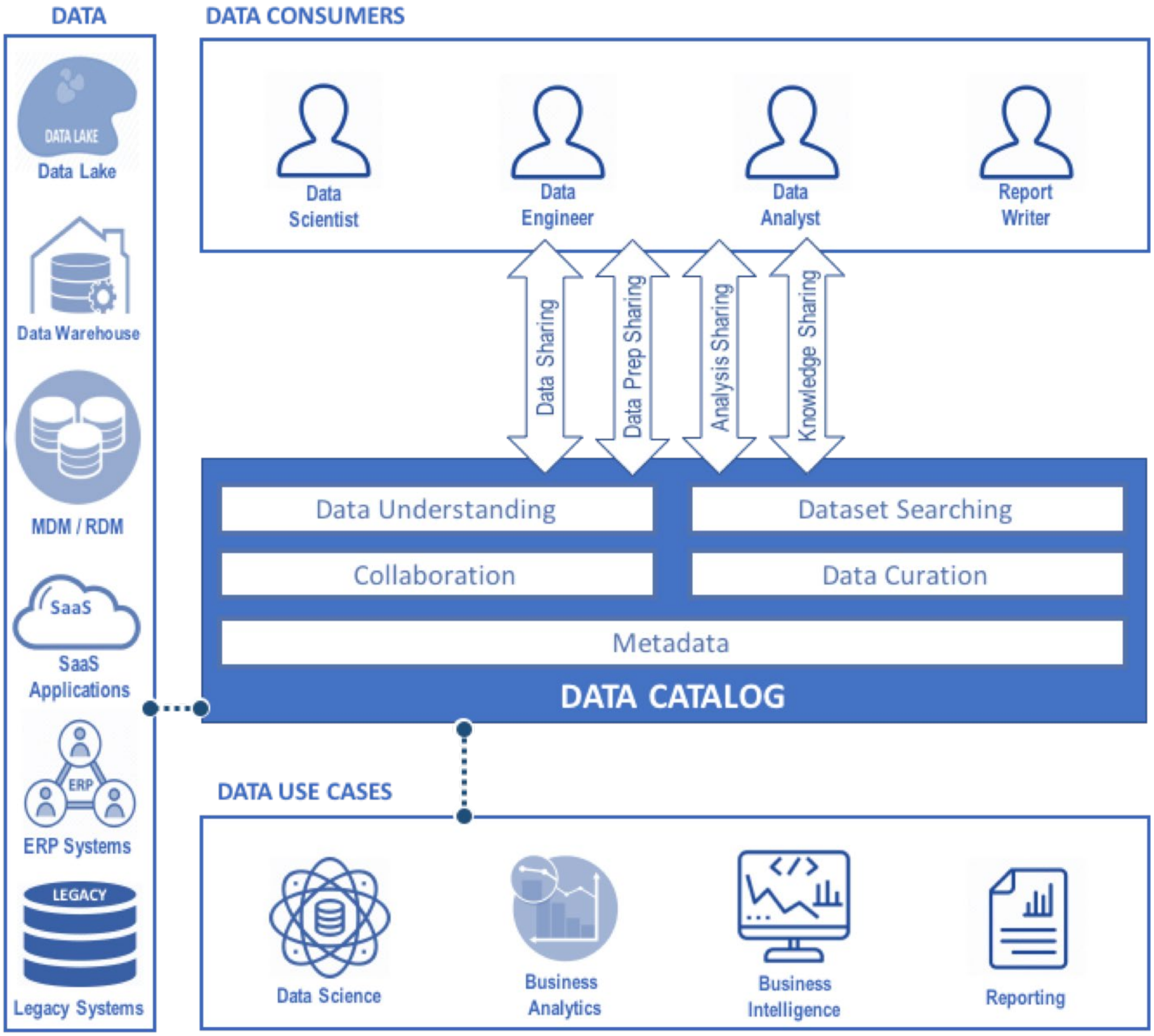


Figure 5. Support for All Users and Collaboration Among Users

7. **Support data quality, data protection, and regulatory compliance.** Managing and mitigating risk is a core function of data management and modern data warehousing. Data quality is at risk when quality deficient data is used for analysis and reporting. Poor quality data harms trust in the data, creates

potential for misinformation, and affects the quality of decision making. Data profiling and algorithmic discovery of data defects and conflicts help mitigate risk of poor data quality. Protecting personally identifying information (PII) and privacy-sensitive data is also an important aspect of data risk management. A modern data warehouse must be able to detect, locate, and classify sensitive data and to protect that data from unauthorized access. In addition to privacy and PII, the data warehouse must remediate risk of non-compliance across the regulatory spectrum including GDPR and a multitude of industry-specific regulations. Compliance risk mitigation is especially important in highly regulated industries such as financial services, healthcare, pharmaceuticals, and energy.

8. **Support a variety of big data processing engines.** There is a multitude of technology choices for processing big data, and the options continue to evolve as open source innovation continues. Many organizations operate multiple processors, in part to optimize the platform for specific data and applications and also because it is impractical when adopting a new technology to go back and convert everything built in the past. A modern data warehouse must support several processing engines and adapt as new technologies emerge. Today's data warehouses should be compatible with the processing engines that many consider to be the top five processing frameworks for big data (see figure 6) including Hadoop, Spark, Flink, Storm, and Samza. Each engine is optimized for particular applications, so limiting to a single processing engine will limit adaptability of the data warehouse.

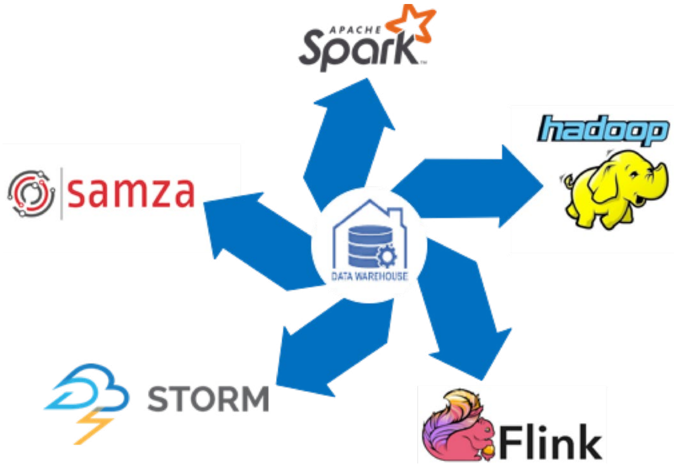


Figure 6. Top 5 Big Data Processing Engines

- Hadoop with MapReduce was the first of the big data processing engines and is still widely used. It works well when data can be processed in batch and processing can be distributed across a cluster.

- Spark is a more recent and more adaptable processing framework than MapReduce and has been adopted widely as the replacement for MapReduce. Spark, which does not have its own distributed storage layer, can be operated within the Hadoop ecosystem and take advantage of HDFS.

- Flink is a stream processing engine capable of batch processing but optimized for streaming and real-time processing of data.
- Storm is a stream processor coupled with a real-time distributed compute engine, making it highly efficient for real-time analytics and machine learning.
- Samza is a distributed stream processing engine that is built on Kafka for messaging and YARN for cluster resource management.

9. **Support the entire data management supply chain.** The processes of data management are much more complex and comprehensive in the age of big data and data lakes than what we experienced when our legacy data warehouses were designed and built. The simple chain of extract → transform → load → publish is no longer sufficient. A modern data management supply chain includes processes for data ingestion, data stream processing, data integration, data enrichment, data preparation, definition and cataloging, mapping of data relationships, data protection, and data delivery. Figure 7 illustrates Informatica's view of a modern data management supply chain.

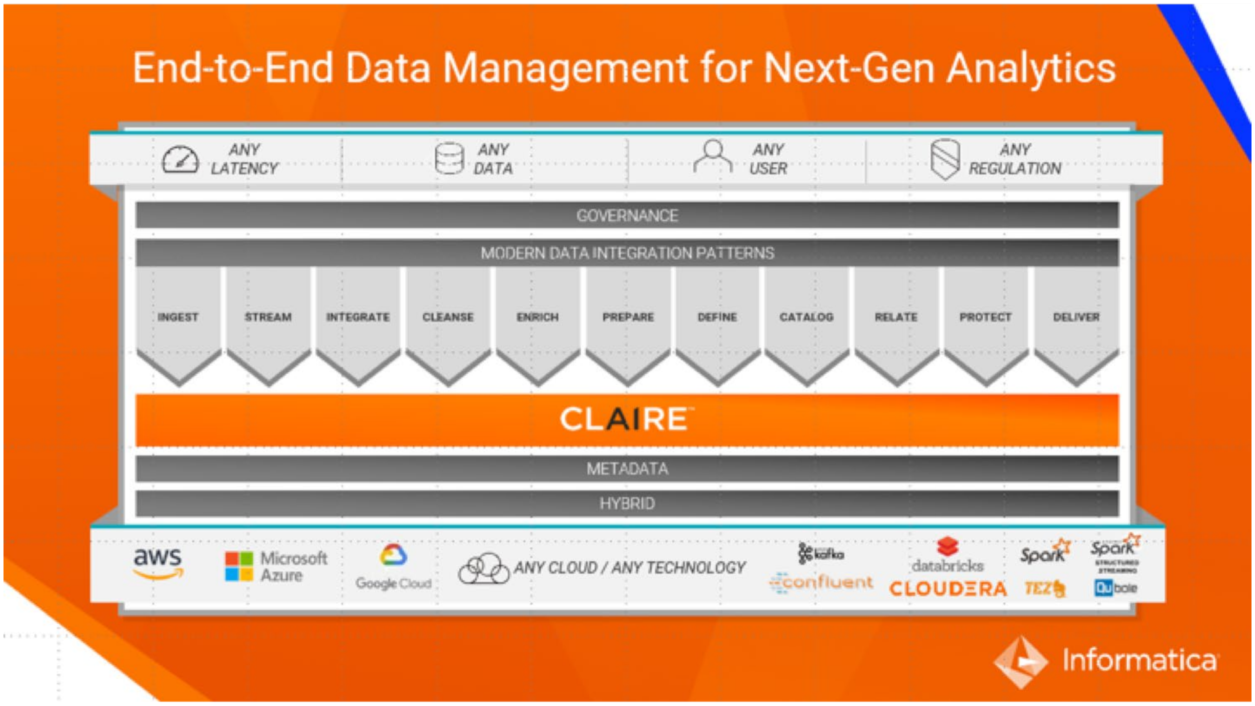


Figure 7. Informatica and the Data Management Supply Chain

10. **Apply AI/ML across the data management supply chain.** It is impractical to attempt data management today without the help of artificial intelligence (AI) and machine learning (ML). Manual data discovery, tagging, matching, mapping, and description simply isn't achievable with the volume, variety, and speed of data. Looking across the supply chain (see figure 7) there are abundant opportunities for algorithms and agents to assist with data management:

    - At data ingestion, the ability to detect and adapt to schema changes minimizes disruption of data pipelines.
    - Data stream processing is optimized and accelerated with algorithmic event parsing.
    - Data integration using smart data transformations and intelligent blending is important when integrating data without shared keys. Blending your internal customer data, for example, with external data that doesn't use your customer ID number can be especially challenging without AI recommendations for matching criteria.
    - Data enrichment may leverage AI to facilitate data cleansing functions as well as discovery of opportunities for data enrichment such as automated geocoding of data with physical address or location information.
    - Data preparation benefits from automation and from recommendations for preparation operations. Masking privacy-sensitive data, for example, is a repeatable preparation step that can be automated through AI and continuously refined through ML.
    - The benefits for data definition, data governance, and data cataloging are immense. Using algorithms to crawl data sources, infer semantics, discover and tag sensitive data, derive metadata, and aid in data curation is an essential part of data management, automating work that is too large in scope and volume to be done manually.
    - Algorithmic discovery of relationships among datasets and intelligent mapping of those relationships enhances data integration, increases data value, aids in data analysis, and simplifies data preparation and blending.
    - Intelligence capabilities to discover, tag, and protect PII, privacy sensitive data, compliance sensitive data, and security sensitive data is an important part of managing and remediating data risks.

- o Data delivery becomes a complex stage of the supply chain when working with many data sources, many use cases, and many data users. AI/ML is useful to build smart data pipelines and to orchestrate the execution of those pipelines.

To summarize, we've now discussed 10 must-have goals and benefits of DWM: use of cloud technologies, support for multi-cloud and hybrid environments, support for all types of data, support for all data latencies, support for all data users, support for collaboration among data users, data risk mitigation and remediation, support for multiple processing engines, support across the data management supply chain, and smart data management with AI and ML.

Data Warehouse Modernization Must Haves

1. Leverage cloud technologies
2. Support multi-cloud and cloud-hybrid
3. Support semi-structured and unstructured data
4. Support batch, real-time and streaming data
5. Support data scientists, analysts, and engineers
6. Foster collaboration among data users
7. Risk mitigation and remediation
8. Support for multiple processing engines
9. Support across the data management supply chain
10. Smart data management with AI ad ML

Four Architectural Patterns for Modernizing Data Warehouses

Data warehousing and the data lake need to work together as complementary components of cohesive data management architecture. There is no one-size-fits-all solution for data management architecture. Every data warehouse is unique, thus every modernization plan is unique. There are, however, several architectural patterns for modernization that help to shift from data warehouse and data lake silos to cohesion and compatibility between data lakes and data warehouses. Use these patterns individually, in combination, or as mix-and-match for multiple warehouses to develop a modernization plan and drive next-generation analytics and AI/ML projects.

Architectural Frameworks

Data Warehousing Outside the Data Lake

This variation treats the data lake and the warehouse as separate data stores without overlap. The data lake is the landing zone for all incoming data, and warehouse ETL draws data directly from the lake. (See figure 8.) The data lake's landing zone serves as warehouse data staging. Sharing a common landing zone for all incoming data reduces redundancy, retains raw data as received, and supports fully traceable data lineage.

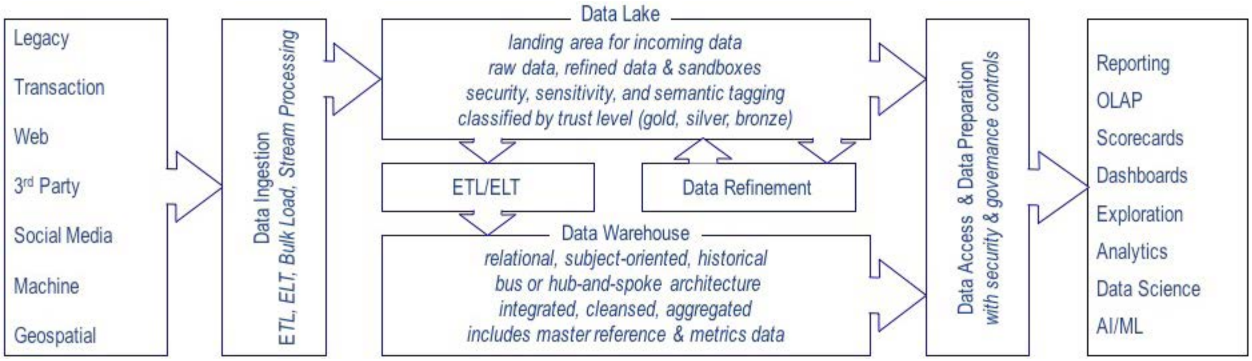

Legacy
Transaction
Web
3rd Party
Social Media
Machine
Geospatial
Data Ingestion
ETL, ELT, Bulk Load, Stream Processing
Data Lake
landing area for incoming data
raw data, refined data & sandboxes
security, sensitivity, and semantic tagging
classified by trust level (gold, silver, bronze)
ETL/ELT
Data Refinement
Data Warehouse
relational, subject-oriented, historical
bus or hub-and-spoke architecture
integrated, cleansed, aggregated
includes master reference & metrics data
Data Access & Data Preparation
with security & governance controls
Reporting
OLAP
Scorecards
Dashboards
Exploration
Analytics
Data Science
AI/ML

Figure 8. Data Warehousing Outside the Data Lake

Data Warehousing Inside the Data Lake

This framework positions the warehouse as part of the data lake. (See figure 9.) The warehouse may acquire data from a raw data zone (data staging) and from a refined data zone where some cleansing and transformation work has already been performed.

It may be especially desirable to position a data warehouse as a subset of the data lake when that warehouse is expected to have a long lifespan with a significant number of users who need to work with raw data, refined data, and integrated and historical warehouse data.

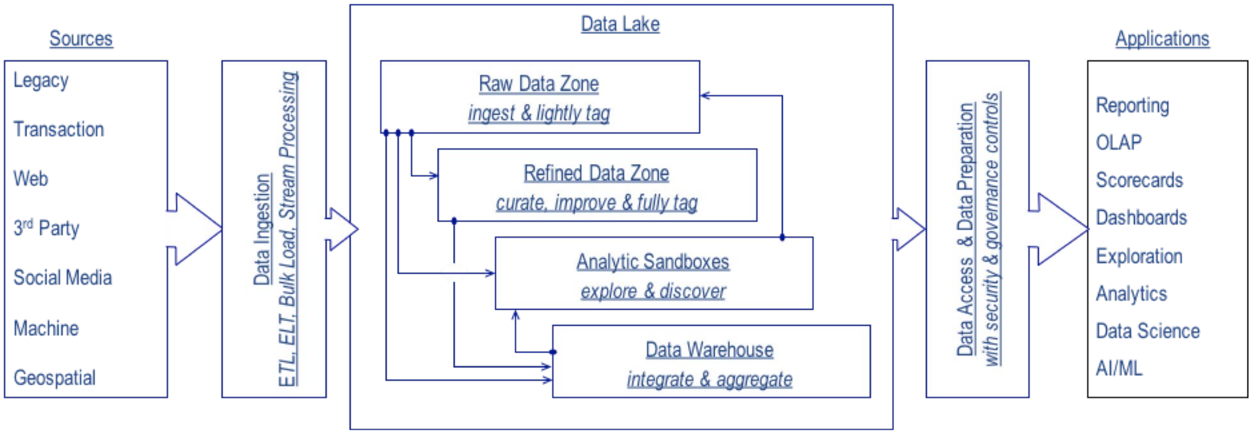


Figure 9. Data Warehousing Inside the Data Lake

Data Warehousing In Front of the Data Lake

In this variation (see figure 10) one or more data warehouses continue to operate independently, but they also become sources for data ingested into the data lake. The modernization advantage here is limited because the data warehouses remain unchanged. Pushing warehouse data to the data lake creates an additional copy of the data, but it also eliminates the silo effect of multiple data warehouses and the data existing separately and in isolation. Although advantages are limited, complexity and effort are relatively small and there is no visible impact to data warehouse users. This may be a practical first step of a multi-phase modernization process.

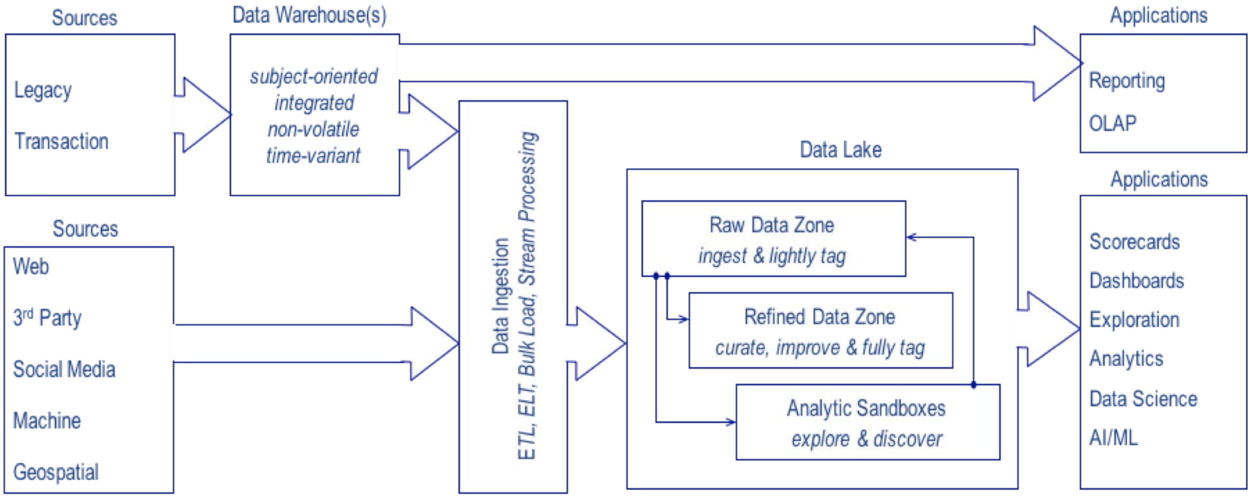


Figure 10. Data Warehousing In Front of the Data Lake

Data Warehouse and Data Lake Inside/Outside Hybrid

With multiple data warehouses, it can be practical to implement a hybrid (see figure 11) where warehouses with heavy analytics usage and overlap with other data lake contents are positioned inside the data lake while those with limited user base and that are primarily used for inquiry and reporting remain outside the data lake.

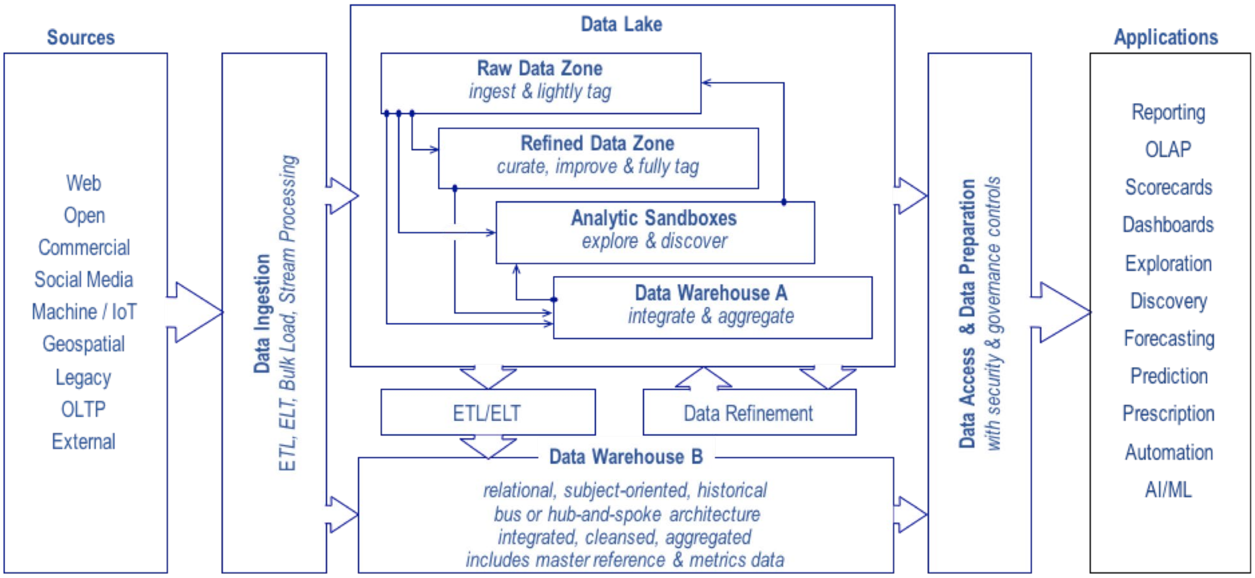


Figure 11. Data Warehouses Inside and Outside the Data Lake

Architectural Decisions

The architectural frameworks shown here are concepts, not prescriptions. Choose among them or blend and adapt them as needed to best fit your data management needs. Consider the number of data warehouses needing modernization, the maturity and stability of your data lake, and your organization's capacity for architectural change. Big changes may offer big benefits, but they require a lot of work. Make architectural modernization practical with an incremental approach. First create the vision of your future architecture, then build the plan to get there one step at a time while taking into consideration the 10 must-haves of data warehouse modernization discussed earlier in this guide

Cloud Platforms for Data Warehouse Modernization

Data warehousing in the cloud has become popular as companies are challenged with growing data volumes, higher service-level expectations, and the need to integrate structured warehouse data with unstructured data in a data lake. The movement to SaaS for enterprise applications also makes cloud data warehousing an inviting option. Cloud data warehousing responds to many legacy data warehouse challenges, offering a targeted and direct response to need for scalability, elasticity, managed infrastructure, cost savings, processing speed, faster deployments, ease of disaster recovery, and improved security and governance capabilities. Less direct but equally important benefits include ready access to technologies designed for non-relational and unstructured data, improved adaptability and agility through instant infrastructure, and reduced dependency on in-house data centers.

Migrating a Data Warehouse to the Cloud

Migrating an existing data warehouse to a cloud platform offers substantial benefits and is a practical step to modernization, but it is not quick or easy. Tactically and technically, data warehouse migration is a challenging multi-step process to move many different warehousing components. (See figure 12.)

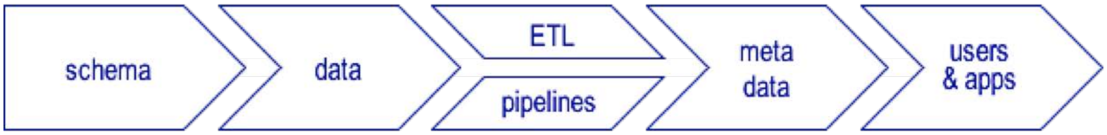


Figure 12. Migrating a Data Warehouse to the Cloud

Migrating a data warehouse to the cloud requires all of the following activities:

- **Migrating Schema.** Before moving warehouse data, you'll need to migrate table structures and specifications, possibly with structural and indexing changes.
- **Migrating Data.** Moving very large volumes of data can be process-intensive, network-intensive, and time-consuming. Don't underestimate resources needed to move the data. If you're transforming data as part of migration, decide whether to transform in stream or pre-process and then migrate.
- **Migrating ETL.** Moving data may be the easy part compared to migrating ETL processes. You may need to change the code base to optimize for platform performance, and change data transformations to sync with data restructuring. This is also an opportunity to reduce data latency.

- **Rebuilding Data Pipelines.** With any substantive change to data flow or data transformation, rebuilding data pipelines may be a better choice than migrating existing ETL.
- **Migrating Metadata.** Source-to-target metadata is a crucial part of managing a data warehouse, knowing data lineage, and tracing and troubleshooting. Consider how you'll move metadata to the cloud platform and how to maintain continuous data lineage that blends pre- and post-migration data movement.
- **Migrating Users and Applications.** The final step in the process is migrating users and applications to the new cloud data warehouse with little or no interruption of business operations. You'll need to migrate security and access authorizations, reconnect BI and analytics tools, and openly communicate with stakeholders throughout.

Common Cloud Platforms

The three common cloud platforms for data warehouse migration are Amazon Web Services (AWS), Microsoft Azure, and Google Cloud. Each works well in conjunction with Informatica CLAIRE—Informatica's intelligent, metadata-driven data management engine—as shown in the following reference architectures. CLAIRE provides support across the entire data lifecycle from data acquisition to data consumption with intermediate steps for data ingestion, preparation, cataloging, security, governance, and access.

Data Warehousing with AWS and Informatica

Figure 13 illustrates a reference architecture for data warehousing with AWS and Informatica CLAIRE.

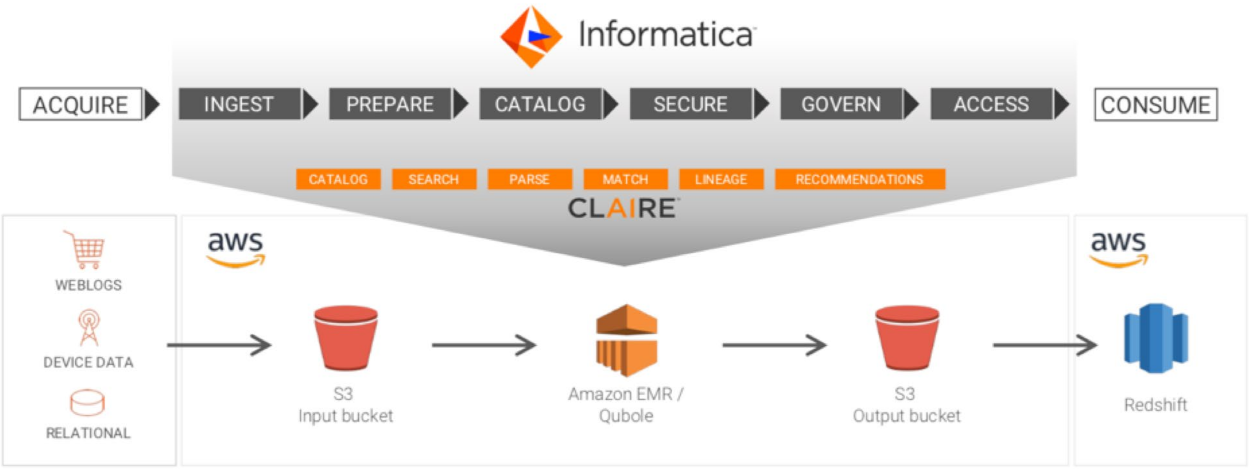


Figure 13. Reference Architecture for Informatica and AWS

Data Warehousing with Microsoft Azure and Informatica

Figure 14 illustrates a reference architecture for data warehousing with Microsoft Azure and Informatica CLAIRE.

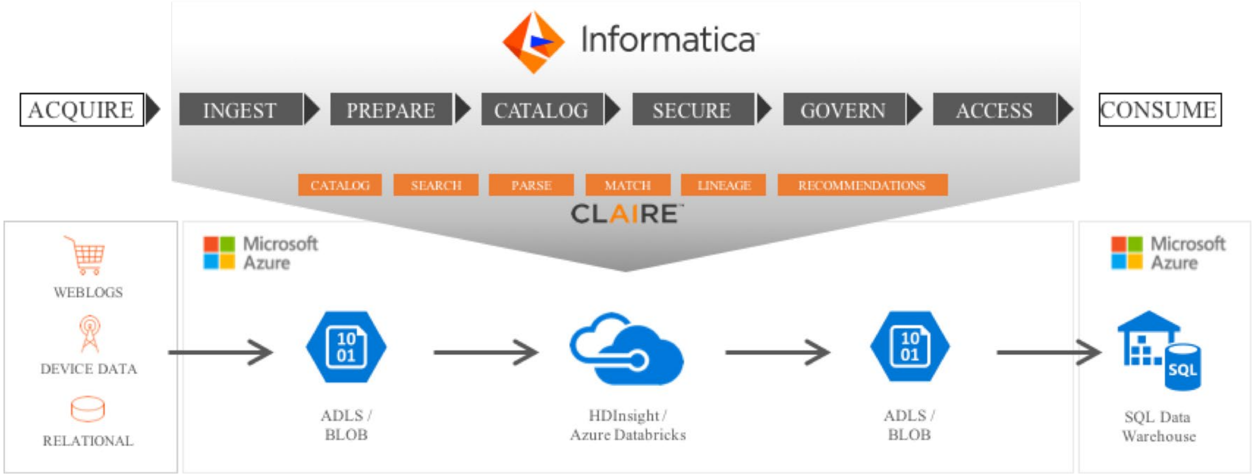


Figure 14. Reference Architecture for Informatica and Microsoft Azure

Data Warehousing with Google Cloud and Informatica

Figure 15 illustrates a reference architecture for data warehousing with Google Cloud and Informatica CLAIRE.

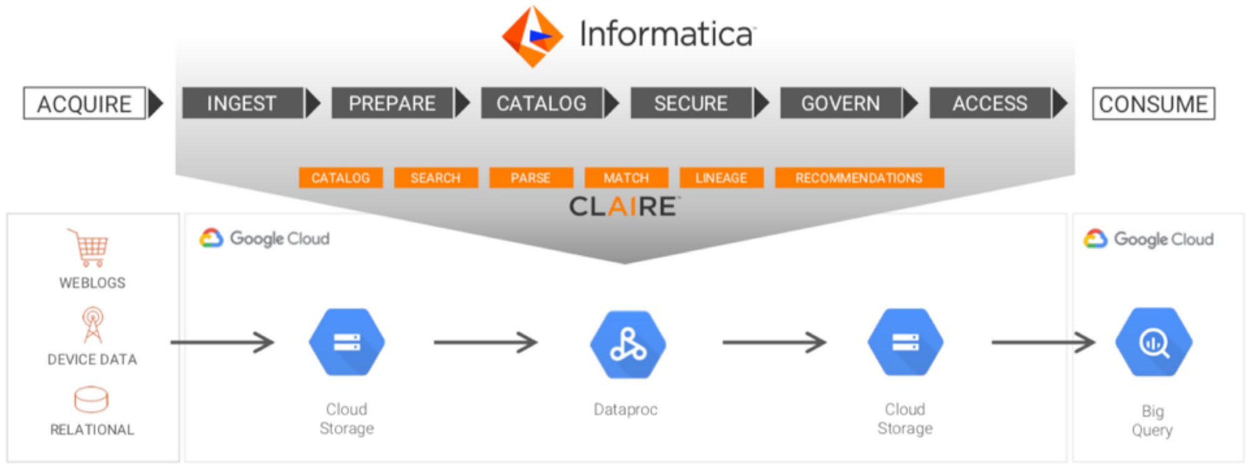


Figure 15. Reference Architecture for Informatica and Google Cloud

Data Warehouse Modernization Use Cases

We have already established that data warehousing continues to be an important part of modern data management. Data warehouses meet the information needs of people and continue to provide value. Many people use them, depend on them, and don't want them to be replaced with a data lake. Data lakes serve analytics and big data needs well. They offer a rich source of data for data scientists and self-service data consumers. But not all data and information workers want to become self-service consumers. Self-service analytics does not replace data warehousing; it extends and complements. Data lakes and data warehouses work together to provide data in a variety of forms from raw data to integrated and aggregated data. They must be designed and managed such that each adds value for the other and they do not exist as isolated data silos. Published data (warehousing) and ad hoc data (self-service) work together to meet a broad spectrum of information needs.

People continue to need well-integrated, systematically cleansed, easy-to-access data that includes time-variant history.

Companies continue to operate data warehouses because they are needed. Business processes and information workers depend on warehouse data and information on a daily basis. Many people—perhaps the majority—continue to need well-integrated, systematically cleansed, easy-to-access relational data that includes a large body of time-variant history. They want to meet routine information needs with data that is prepared and published with those needs in mind.

The variety of use cases that exist in a data-driven business are many, and there is no one-size-fits-all data organization that is optimized for all users and uses. Data warehouses and data lakes should work together to provide a variety of data to meet all uses cases. The following pages illustrated several common data uses cases where coexistence of data warehouses and data lakes is central to meeting the data and information needs. You probably have several of these use cases in your organization and it is likely that you'll have others not shown here. Each use case is illustrated with a corresponding reference architecture. Use these as a baseline from which you can adjust to adapt to your unique data and use case characteristics.

Streaming Analytics

Data streams are among the most challenging of big data sources. Data arrives in real time from machines, sensors, or other IoT connected devices. If you use RFID tagging, GPS enabled devices, or robotics you are certain to have streaming data. It must be captured and/or analyzed as it arrives. Streaming data is acquired by connecting to the stream. Individual events are parsed from data stream upon ingestion, sometimes filter to include only events of interest. Event data is typically collected as raw data in a data lake. Events may be analyzed in real time to send alerts when a measurement exceeds a threshold or otherwise indicates need for immediate attention. Sometimes event data may flow directly to dashboards for real time monitoring. Deeper analysis and reporting usually require that the data is enriched to add context. Typical machine and sensor data is sparse, including only the machine/sensor id, a measurement value, and a date/time stamp. Adding context such as machine or sensor attributes relies on persistent reference data that is often found in a data warehouse. The data warehouse may also collect time-variant history from a data stream to support time-series and trend analysis.

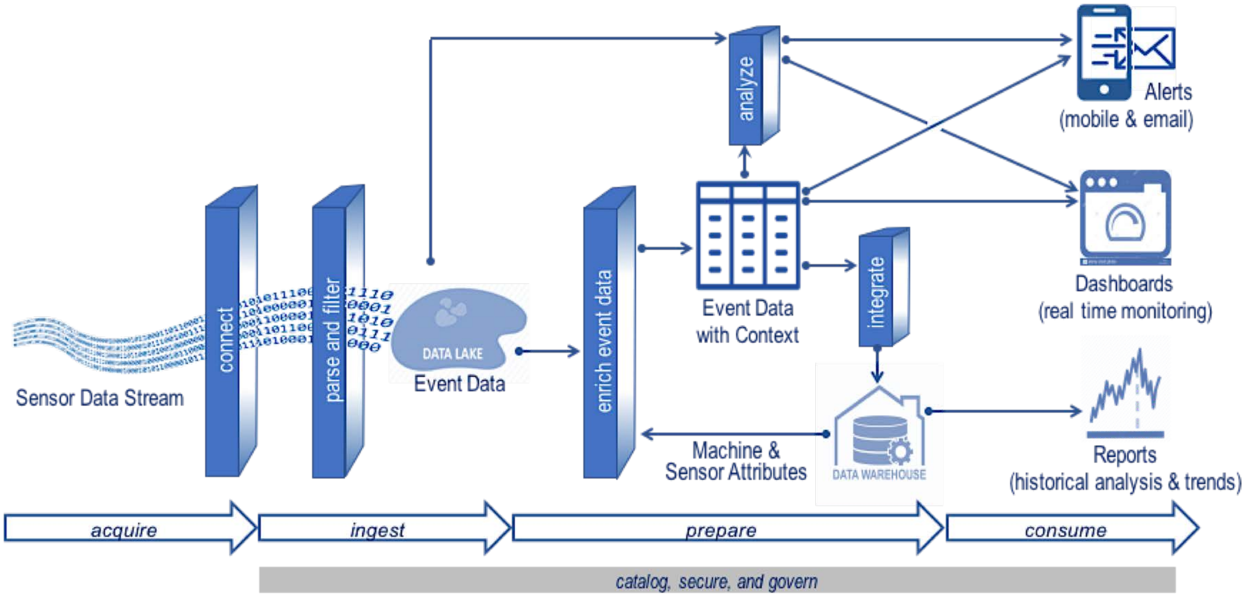


Figure 16. Streaming Analytics Reference Architecture

Self-Service Analytics

Self-service analytics is a ubiquitous use case. Every organization with Tableau, Qlik, Power BI, or similar tools has self-service data analysts who are continuously faced with the challenge of finding and understanding data. Data scientists face many of the same challenges, needing to find the right data for their modeling efforts. It is often said that these analysts and scientists spend 80% of their time finding and preparing data, and only 20% analyzing and finding insights. They often struggle to determine where to go for data—lake, warehouse, or elsewhere. When data is cataloged upon ingestion and prepared, much of the struggle is eliminated and the 80/20 rule is reversed, with 80% of time spent on analysis and insights. In a smart data ecosystem, much of data cataloging and data preparation is automated with AI and machine learning discovering data characteristics, inferring semantics, tagging sensitive data, and making data searchable. This also facilitates collaboration, allowing users to share data knowledge, data preparation operations, and even data analysis.

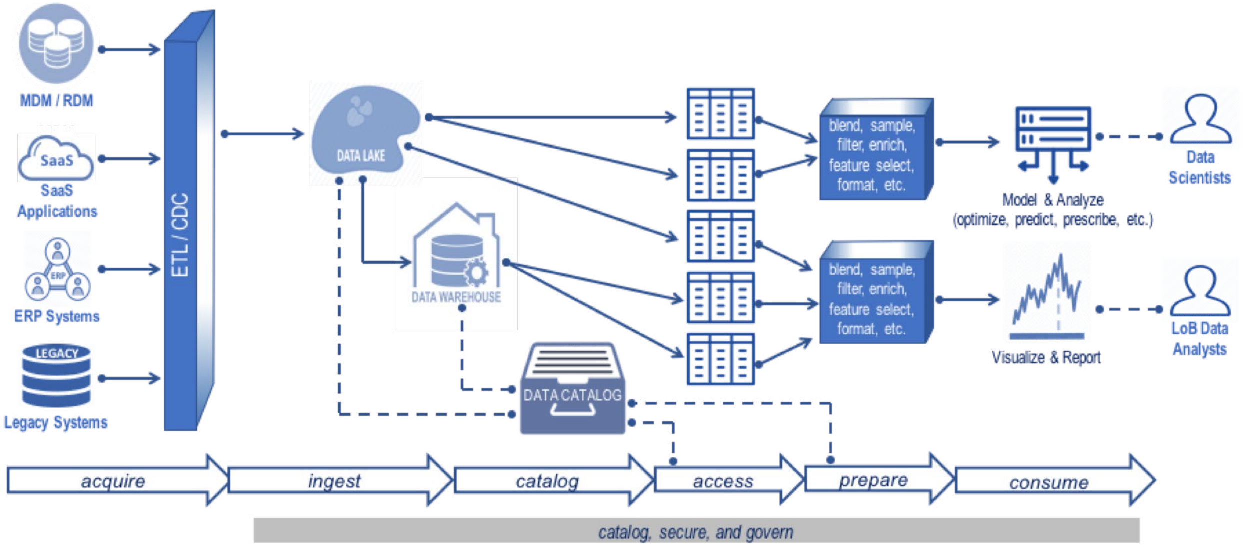


Figure 17. Self-Service Analytics Reference Architecture

Hybrid Environment for Analytics

The challenges of finding data are compounded when data exists in a complex environment that includes multiple cloud platforms as well as on-premises data—a reality for most organizations today. Perhaps we should stop saying “in the cloud” and instead talk about “in the clouds” as most of us work in a multi-cloud environment of SaaS applications, cloud hosted ERP systems, and cloud data lakes. Yet we also have on-premises data sources and frequently on-premises data warehouses. Working with data spread across multiple platforms has unique challenges for finding, accessing, and blending data. The data catalog fills a critical role here with all of the benefits described for self-service analytics—finding data, understanding data, preparing data, and collaborating when working with data. In a complex multi-cloud/hybrid data ecosystem, data prep together with a data catalog helps users to find and enrich data regardless of deployment platform, and without the need to know where the data is hosted.

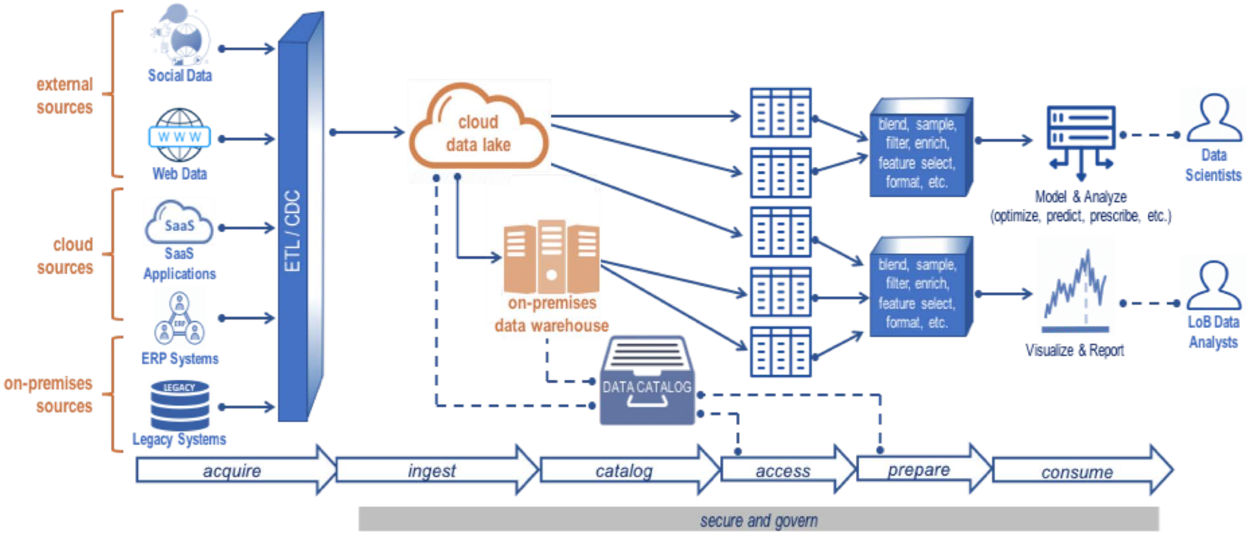


Figure 18. Hybrid Environment for Analytics Reference Architecture

Big Data/Hybrid Integration Hub

Sometimes the data environment is so large and complex that a data integration hub is the right solution. When you have an exceptionally large and diverse number of on-premises and cloud data sources combined with many and varied users and use cases, data integration driven by individual sources or uses is impractical. A cloud-based data integration hub brings data together in a single location to harmonize without proliferating redundant copies of data. A robust data hub includes features for data storage, harmonization, indexing, processing, governance, metadata, search, and exploration. Note in this reference architecture that the data lake and data warehouse still exist, taking on new roles as sources of data for the integration hub.

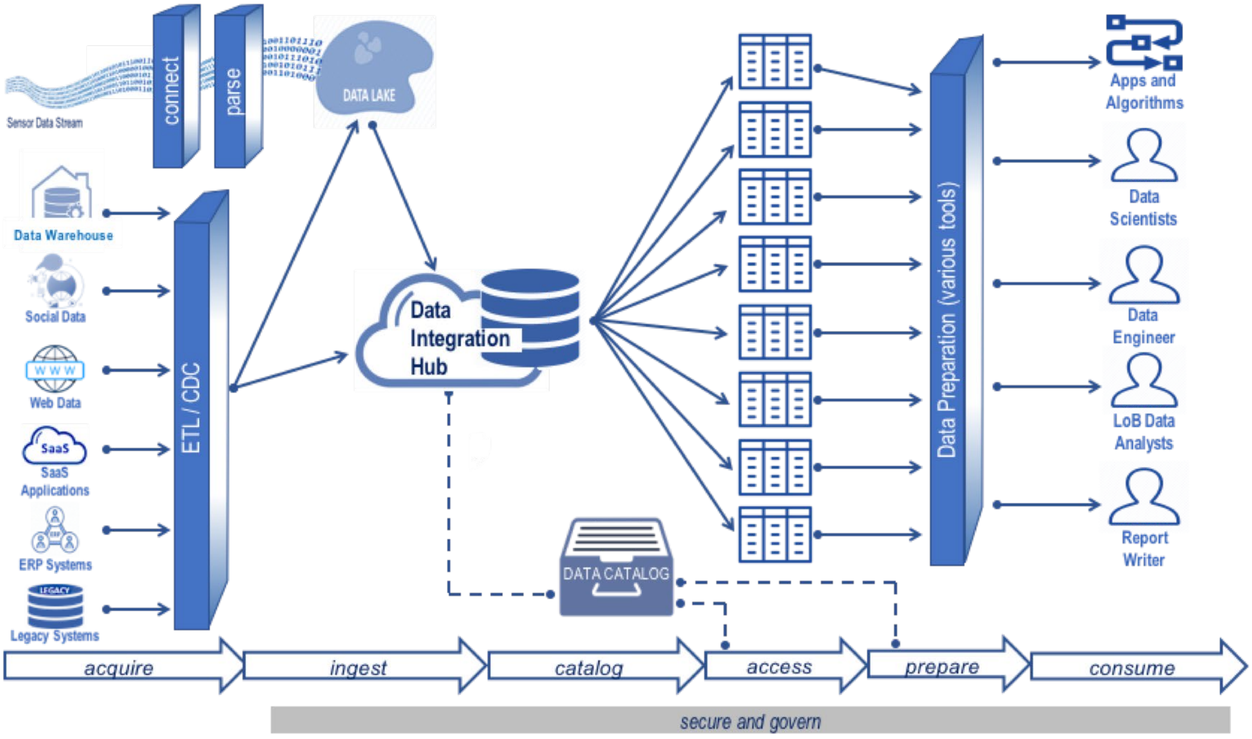


Figure 19. Big Data Integration Hub Use Case

Advanced Analytics, AI, and Machine Learning

Advanced analytics (predictive and prescriptive), artificial intelligence (AI), and machine learning (ML) are at the cutting edge of modern data use cases. Algorithm-based data applications ranging from decision automation to robotics and autonomous devices offer great opportunities for digital transformation of business. But they may also bring high risk and potential for negative consequences. Data quality is a critical consideration for these kinds of applications. Consider, for example, the risks inherent in poor data quality for diagnostic and prescriptive decision automation in healthcare. Similarly, a machine learning application in social sciences working with quality deficient data would learn incorrectly, produce biased algorithms, and potentially disrupt lives. Data quality is critical for prediction, prescription, automation, AI and ML. Data quality assurance and data cleansing must be included as part of data preparation work.

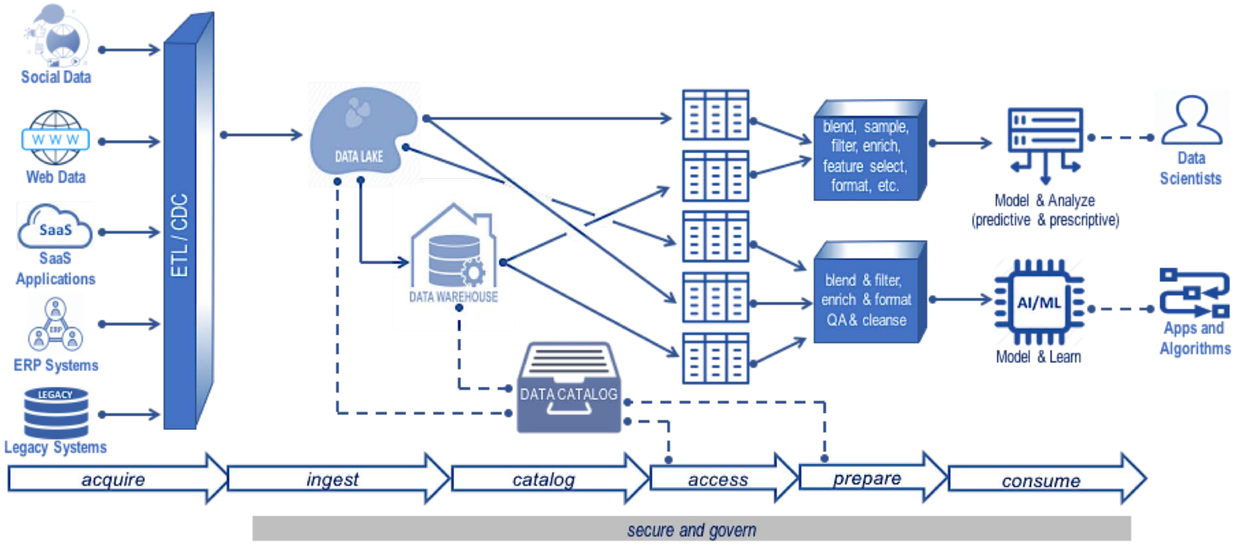


Figure 20. Advanced Analytics and AI/ML Reference Architecture

Getting Started with Modernization

Data warehouse modernization is a journey, not an event. Done well it is a planned and incremental step-by-step process of moving from warehousing of the past to the future of data warehousing. Plan your journey and navigate the course with these tips in mind:

- Start with an inventory of data warehouses (you probably have more than one) as well as other data sources. Know the purpose of each and who uses them. Identify overlaps and redundancies across multiple warehouses.
- Assess your current state of data warehousing. Define your needs and challenges and prioritize them to know which are most pressing and need earliest attention.
- Identify current high-priority use cases, then look ahead three to five years to consider probable future use cases.
- Define your future state of data warehousing. Define and describe your goals for modernization with enough clarity to know when the goals have been achieved.

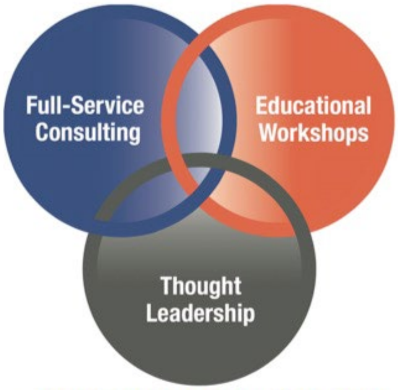
- If not already done, map out your existing data management architecture from data sources, through warehouses and lakes, to delivery of data to consumers.
- Rethink data management architecture for maximum cohesion with coexistence of data lake and data warehouse.
- Choose the modernization patterns that are best suited to your goals. Don't hesitate to mix and match patterns. You might, for example, want to use data warehouse automation to reverse engineer a legacy data warehouse and then migrate that warehouse to the cloud. Or you might migrate a high use data warehouse to the cloud and federate other warehouses to break down the silos.
- Plan for the future, not for today. Look ahead at least three years. Planning only for today leads to being outdated before you're fully implemented. Be sure to make technology choices with long-term goals in mind. Future-proof your technology investments by selecting technologies that use AI/ML across the entire data management lifecycle from ingestion to consumption.
- Execute one step at a time, not "big bang." Then repeat the entire process. After each step your current state will be different, your priorities may have changed, and you may want to refine your future state thinking. And, of course, the technology will continue to evolve.

About Eckerson Group

Wayne Eckerson, a globally known author, speaker, and advisor, formed Eckerson Group to help organizations get more value from data and analytics. His goal is to provide organizations with a cocoon of support during every step of their data journeys.

Today, Eckerson Group helps organizations in three ways:

- **Our thought leaders** publish practical, compelling content that keeps you abreast of the latest trends, techniques, and tools in the data analytics field.
- **Our consultants** listen carefully, think deeply, and craft tailored solutions that translate your business requirements into compelling strategies and solutions.
- **Our educators** share best practices in more than **30 onsite workshops** that align your team around industry frameworks.



Get More Value From Your Data

Unlike other firms, Eckerson Group focuses solely on data analytics. Our experts each have more than 25+ years of experience in the field. They specialize in every facet of data analytics—from data architecture and data governance to business intelligence and artificial intelligence. Their primary mission is to help you get more value from data and analytics by sharing their hard-won lessons with you.

Our clients say we are hard-working, insightful, and humble. We take the compliment! It all stems from our love of data and desire to help you get more value from analytics—we see ourselves as a family of continuous learners, interpreting the world of data and analytics for you and others.

Get more value from your data. Put an expert on your side.
Learn what Eckerson Group can do for you!

About Informatica

Informatica, the leader in enterprise cloud data management, provides an AI-driven, microservices-based Intelligent Data Platform with solutions that are purpose-built for Microsoft Azure. These solutions are designed to accelerate the migration to Azure by automating your data integration development lifecycle, including connectivity, development, deployment, and management.

informatica.com