Increase Agility and Reduce Costs with a Logical Data Warehouse

February 2014

Table of Contents

Summary

A Logical Data Warehouse solves the problem of consolidating critical data scattered across silos, providing a comprehensive, actionable view of data assets. Coined by Gartner, the “Logical Data Warehouse” is also referred to as ‘data virtualization,’ ‘data layer,’ ‘data cloud,’ ‘data lake’, ‘data fabric’, and many other terms referred to as data virtualization, data layer, data cloud, data lake, data fabric, and many other terms depending on the industry and environment.

No matter what term is used, a Logical Data Warehouse solution powered by MarkLogic is an active, searchable enterprise data layer that presents a unified view of multi-structured data across organizational silos. It provides a set of services to aid in information management, provisioning of tools, and secure information sharing. Organizations deploy Logical Data Warehouses to greatly reduce implementation time and costs associated with new information products and services, to support secure information sharing across different user communities, and to "future proof" their technical infrastructure against changes to data sources.

MarkLogic combines best-of-breed technology across database management systems, application servers, and search engines. Some key benefits of MarkLogic in a Logical Data Warehouse infrastructure include improved analytics and agility, higher performance and scalability, lower total cost of ownership (TCO), and faster time to value than systems based on traditional technologies such as relational databases. The MarkLogic platform is architected to efficiently handle the high volumes of varied and complex types of information found in an enterprise environment. With MarkLogic, organizations get a comprehensive, actionable view of their mission-critical data in an agile and cost effective architecture, to make better decisions faster with lower TCO.

For more than a decade, MarkLogic has delivered a powerful, agile, and trusted Enterprise NoSQL database platform that enables organizations to turn all data into valuable and actionable information. Organizations around the world rely on MarkLogic's enterprise-grade technology to power the new generation of information applications.

Data Virtualization & the Logical Data Warehouse

What is a Logical Data Warehouse?

A Logical Data Warehouse powered by MarkLogic is an active, searchable 'data layer' or 'data cloud' that presents a unified view of multi-structured and unstructured data across organizational silos. It leverages a metadata-centric data virtualization approach to make data accessible without requiring costly data duplication. The MarkLogic solution includes an Enterprise NoSQL database, web services and tools for connecting with remote data sources and ingesting high volumes of data, and APIs for delivering data to multiple applications. It provides data operations such as search, discovery, analytics, update, enrichment, normalization, transformation, and delivery, while consistently applying security controls and maintaining real-time transactional consistency (ACID-compliant).

Why Deploy a Logical Data Warehouse?

Too often, critical information becomes scattered across silos, frequently in proprietary or unmanaged formats. This typically happens over time due to organizational IT processes, new information sharing partnerships, reorganizations, or mergers and acquisitions. The information is not only of many varied formats, but is also owned by separate divisions or outside organizations, creating challenges for data access and aggregation. In this environment, it is impossible to centralize data into a single repository for new search and analytics applications. Thus, the organization and its end-users pay the price in terms of the time it takes to find and analyze all relevant information to support decision-making. A Logical Data Warehouse can resolve this issue by ensuring that information across an organization and its partners is securely discoverable and usable, that new information is readily incorporated, and that supporting IT costs are kept to a minimum.

In comparison, traditional environments support only limited discovery, often via rigid, custom point-to-point connections – or extensive up-front data modeling and ETL efforts in order to get information into a traditional data warehouse. Such environments are inflexible when confronted with new types of information or new access requirements (Figure 1). Also, systems commonly used today are limited to statically exposing data according to its fixed structure or schema – and even in best practices environments, structured and unstructured data are accessed separately – structured data through an enterprise data warehouse, and unstructured data via an enterprise search engine.

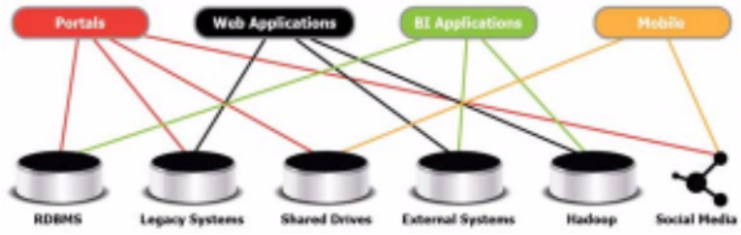
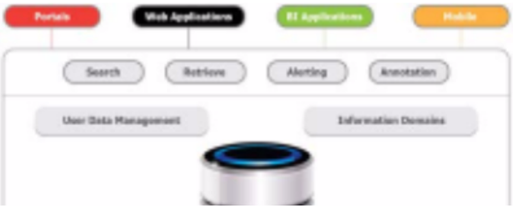


Figure 1. Without a data consolidation layer, discovery is limited or expensive due to custom integrations.

By contrast, a Logical Data Warehouse powered by MarkLogic sits at the center of information flows (Figure 2), which it dynamically manages using metadata properties and policies. It streamlines the management, creation, and dissemination of information for users across the enterprise. Having a central point of control greatly reduces development efforts, breaks down the information silos, and provides secure access while facilitating real-time data sharing and knowledge discovery. Legacy applications integrate with MarkLogic using standards-based services for ingest, search, federation, and alerting. New data sources can be added to the data layer and associated with existing information without costly re-engineering.



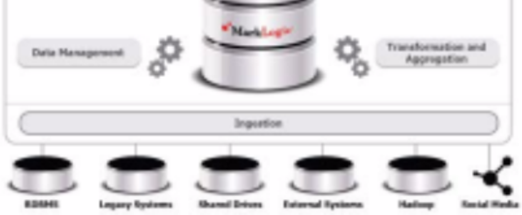


Figure 2. A Logical Data Warehouse provides a unified data view and a central access point for information applications.

A Logical Data Warehouse Based on the MarkLogic Enterprise NoSQL Database Platform

New types and growing volumes of data – combined with insatiable demands from business users for flexible, ad hoc analytics and applications to leverage that data – have pushed the limits of legacy data integration and warehousing approaches to their breaking point. Organizations are suffering from lengthy development cycles, data latency, data quality and governance challenges, limits on business applications, and high costs.

The challenges organizations face in unifying disparate data silos pertain to the volume, velocity, variety, and complexity of data – otherwise known as "Big Data." MarkLogic was specifically architected to handle Big Data, making it an ideal foundation for data consolidation. Our customers benefit from the flexibility of a NoSQL data model, the reliability of an enterprise-grade database, and the precise search and retrieval capabilities of a best-of-breed search engine. From day one, MarkLogic was built to handle complex, real-time transactional data. With hundreds of deployments in information-rich environments, MarkLogic has a proven track record for handling these data challenges. The core capabilities of MarkLogic, in conjunction

with value-added functionality, provide the capabilities today's organizations need to get a comprehensive view of enterprise data.

Key features of MarkLogic for data virtualization and unification include:

- Ingestion Services. Data is loaded into MarkLogic via a standard set of ingestion services that support multiple access protocols (REST, SOAP, etc.). Data can be loaded "as is," without need for data modeling, to enable immediate discovery, and can be subsequently transformed to address application-specific needs. IT teams can use the browser-based MarkLogic Information Studio tool to help with quick and efficient loading of information into MarkLogic. For extremely high volumes of data, the MarkLogic Content Pump (mlcp) helps IT teams realize lower operation and maintenance costs of ongoing data import, and speed time-to-deployment by loading large data sets in parallel.

- Core Logical Data Warehouse / data virtualization layer components:

    - Metadata catalog and repository. The repository stores data that range from basic metadata, to rich, detailed metadata plus the "source" content itself. The repository incorporates semantic models to map data into meaningful views of entities and relationships. These models facilitate both data discovery and analytics and can be predefined (e.g., by deploying industry standard ontologies) or incorporated on an ad hoc basis via user tagging. Some organizations also choose to completely move certain critical data out of obsolete legacy systems and into MarkLogic, which allows them to turn off these older, more expensive systems.

  - Transformation and Aggregation. Data consolidation solutions powered by MarkLogic can provide value-added data enrichment for improved search and analytics. This includes data conversions such as date format standardization, field name standardization, entity extraction, and data amalgamation.

  - Data management. MarkLogic incorporates rules for data retention, automatic purging, and load balancing.

  - User data management. These services allow users to create, retrieve, update, and delete named saved searches and alert criteria.

  - Information domains. MarkLogic data consolidation solutions are generally oriented around “information domains,” which not only define views for different user communities but also control access to data (who can see and do what). This makes it easier for the organization to securely deliver the appropriate data to that community, e.g., by subject matter, geographic region, or national security classification.

- Application Services and Tools:

  - MarkLogic provides a number of application services and APIs, with support for Java, REST, JSON, and XQuery:

    - Search. MarkLogic provides full-text, faceted, temporal, and geospatial search for all documents, values, and triples while adhering to data access controls. The MarkLogic Enterprise NoSQL database platform incorporates both role- and attribute-based security, ensuring users can only find information they are authorized to see.

    - Retrieval. Once desired information is located by the user, MarkLogic can retrieve the items from their source locations if the user has proper authorization.

    - Alerting. MarkLogic provides the ability for users to save high-performance, real-time queries for searching new and updated data.

    - Annotation. MarkLogic provides the ability to associate entities to documents and add simple “tags” or detailed notes to documents. This user-controlled annotation is in addition to any data mapping or tagging that occurs upon ingest, or any semantic services that are working upon the data stored in the repository.

  - Another MarkLogic tool to help speed application development is Application Builder, which creates HTML5 search applications based on the MarkLogic REST API. Developers can rapidly prototype new applications or even generate production-ready, baseline applications without writing any code. Application Builder includes Visualization Widgets to create graphical representations of the data in the repository.

All of these features are built on top of the core MarkLogic technology, which includes enterprise-ready capabilities such as high availability, disaster recovery, role- and attribute-based security, elastic tiered storage, and a full suite of monitoring and management tools.

Benefits of Using MarkLogic for a Logical Data Warehouse

With MarkLogic, organizations can get a comprehensive, complete view of enterprise data with the agility to quickly address changing requirements, with lower cost when compared to traditional enterprise data warehouses such as those based on an RDBMS.

Comprehensive View of All Critical Data Assets

Organizations can make better decisions when they have a broad, unified view of all their data. Unlike many traditional data integration providers, MarkLogic supports comprehensive and flexible data virtualization that includes support for the volume, variety, velocity, and complexity of today’s information sources.

A key element of the MarkLogic platform is an open, schema-agnostic data model to leverage multi-structured and unstructured information. This is useful for integrating data from disparate silos in distinct formats, as it eliminates much of the data modeling effort required in traditional systems. In addition, MarkLogic can further enrich content with automatic tagging and/or inline (user-controlled) markup, thus enabling more precise search and analytics.

Agility in Application Development and Analysis

Organizations need to adapt as new requirements arise, and MarkLogic’s Enterprise NoSQL database platform is designed to rapidly handle new data types and analytic requirements. Key features of this agility include:

DATA MODEL FLEXIBILITY FOR ADAPTABILITY AND DECREASED ADMINISTRATIVE LOAD

Neither a fixed schema nor any up-front data modeling is required as MarkLogic is schema-agnostic and loads data “as is.” This lets it efficiently ingest, search, and store metadata and other XML documents. When compared to an RDBMS – and its time-consuming data modeling, planning, and re-formatting requirements – MarkLogic avoids significant administrative and development overhead.

PRECISE, ITERATIVE, REAL-TIME QUERYING AND ALERTING

With out-of-the-box, embedded support for real-time search, MarkLogic ensures information can be searched immediately after it is ingested. It eliminates the indexing latency typically found in applications that rely on separate systems to handle storage and search. MarkLogic provides the search capabilities found in enterprise search engines today, including advanced features such as faceted navigation, autofound in enterprise search engines today, including advanced features such as faceted navigation, auto-suggestion, relevance tuning, and advanced international language processing.

Real-time alerting pushes information to users as soon as it is available. Millions of alerts can be created to enable a variety of immediate actions, including delivery, categorization, enrichment, and transformation. Alerts are optimized for this type of matching, and are far more efficient than systems that periodically run standing queries.

SERVICES-ORIENTED ARCHITECTURE (SOA) FOR PROVISIONING OF INFORMATION SERVICES

MarkLogic is entirely service-enabled and easily extends existing SOA environments by providing data services for data ingestion as well as data delivery to end-user applications such as business intelligence tools. Web services for ingest, search, retrieval, alerts, and annotation are provisioned and secured from policies defined within the metadata catalog.

Cost-Effectiveness

Organizations using MarkLogic get fast return to business value, at lower TCO than competing solutions. As opposed to traditional IT infrastructure initiatives that may take years before providing value, the MarkLogic architecture and focus on mission-critical data allows an organization to realize value in weeks instead of years. Highlights of the MarkLogic data virtualization approach include:

STANDARDIZED APIS FOR LOWER COST APPLICATION DEVELOPMENT

Once data assets have been incorporated into the MarkLogic consolidated data view, they can be readily delivered to a variety of end-user applications either via web services or by using the embedded MarkLogic application services to build applications directly on top of the Logical Data Warehouse. Simple REST interfaces built on MarkLogic offer a low learning curve for building new applications.

SHARED-NOTHING ARCHITECTURE FOR HORIZONTAL, MASSIVE SCALABILITY ON COMMODITY HARDWARE

MarkLogic leverages a "shared-nothing" architecture to scale out a cluster on commodity hardware as data and/or user loads grow. This architecture simplifies the addition of new nodes to a cluster, and efficiently uses hardware by providing linear scaling. Additionally, MarkLogic provides elastic tiered storage that allows organizations appropriately size their infrastructure, and maintain their data in the most economically sensible place, including a mix of HDFS, cloud, SAN, NAS, SSD, or local disk.

OPTIMIZED DATABASE ADMINISTRATION

Organizations gain cost savings by leveraging commodity hardware servers, requiring fewer servers than would be required in a RDBMS-based environment, and dramatically lowering administrative overhead.

Conclusion

The metadata-centric data virtualization approach described in this paper has been implemented by customers in both the public and private sectors to provide access quickly—usually within a few weeks—to data assets located in silos both within and outside their organizations. This helps them to analyze and share information vital to national security, healthcare, asset management, and other core business initiatives. This approach creates a new breed of data warehouse – the Logical Data Warehouse – which provides organizations with significant agility and cost savings when managing data and delivering information applications.

Some of these customers have also decided to extend their Logical Data Warehouse deployments to leverage the unique capabilities of MarkLogic to store, manage, and exploit even more of their mission-critical information and applications. Since MarkLogic is an Enterprise NoSQL database, it can be used to ingest and manage not only metadata but also any other type of multi-structured or unstructured mission-critical Big Data. This allows organizations to consolidate information from expensive legacy silos—such as read-only mainframe applications—that can then be decommissioned, incurring savings from the reduction in operations and maintenance overhead costs. Extensions to the MarkLogic solution are generally implemented in development iterations spanning only a few weeks, quickly and incrementally driving new value to the organization.

About MarkLogic

For more than a decade, MarkLogic has delivered a powerful, agile, and trusted Enterprise NoSQL database platform that enables organizations to turn all data into valuable and actionable information. Organizations around the world rely on MarkLogic's enterprise-grade technology to power the new generation of information applications. MarkLogic is headquartered in Silicon Valley with offices in Washington D.C., New York, London, Frankfurt, Utrecht, and Tokyo. For more information, please visit www.marklogic.com.

999 Skyway Road, Suite 200, San Carlos, CA 94070 ›US: +1 650 655 2300 › INT'L.: +1 877 992 8885
sales@marklogic.com › www.marklogic.com