Creating an Actionable View of Data with Data Virtualization

January 2013
Executive Summary

Data virtualization solves the problem of consolidating critical data scattered across silos, providing a comprehensive, actionable view of data assets. A data virtualization solution powered by MarkLogic is an active, searchable ‘data layer’ or ‘data cloud’ 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 MarkLogic 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 data virtualization 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 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 and trusted next-generation Enterprise NoSQL database that enables organizations to turn all data into valuable and actionable information. Organizations around the world rely on MarkLogic’s enterprise-grade technology to make better decisions faster.
**Data Virtualization**

**What is Data Virtualization?**

A data virtualization solution powered by MarkLogic is an active, searchable ‘data layer’ or ‘data cloud’ that presents a unified view of multi-structured data across organizational silos. It consists of 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.

**Why Use Data Virtualization?**

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. Data virtualization 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. 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.

![Figure 1](image.png)

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

By contrast, a data layer 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 data layer provides a unified data view and a central access point for information applications.

Data Virtualization Based on the MarkLogic Enterprise NoSQL Database

The challenges of unifying disparate data silos are similar to the challenges of handling today’s Big Data. That is, the challenges pertain to the volume, velocity, variety, and complexity of data. MarkLogic was 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 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.
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 documents and entities, while adhering to data access controls. The MarkLogic Enterprise NoSQL database 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 those documents and entities 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, and a full suite of monitoring and management tools.

**Benefits of Using MarkLogic for Data Virtualization**

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 RDBMS-based solutions or traditional data warehouses.
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 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 data model to leverage multi-structured 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 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 which requires time-consuming data modeling, planning, and re-formatting, one can see significant administrative and development overhead is avoided with MarkLogic.

- Precise, iterative, real-time querying and alerting

With 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, 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 data virtualization layer. Simple REST interfaces built on MarkLogic offer a low learning curve for building new applications.

- **Shared-nothing architecture for elastic, 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.

- **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.

**Summary**

The metadata-centric data virtualization approach described 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.

Some of these customers have also decided to extend their data virtualization 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 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 significant 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 Corporation**

Ever since MarkLogic Corporation was founded in 2001, we have focused on building a database and applications that enable our customers to capture more data, and do more with it. We give our customers an unmatched competitive edge through game-changing technology. We’ve set new standards in scalability, enterprise-readiness, time to value, and innovation.

The world is seeing an explosion in data, including user-generated content, machine-generated content, aggregation of highly structured heterogeneous data sources (patient records, insurance claims, mortgage documents, etc.), raw data, social media, log files, and sensor data, among others. Relational databases were not designed to deal with this variety and complexity in real-time.

Some businesses, organizations, and agencies are missing out on a huge opportunity. Our customers are not. The Federal Aviation Administration depends on MarkLogic as the backbone of its Emergency Operations Network. Conde Nast is building new applications in weeks. CQ Roll Call puts up-to-date information about government-in-action at the fingertips of its subscribers. These are just a few MarkLogic customers who are revolutionizing the industry and the public sector with Big Data Applications.

MarkLogic is headquartered in Silicon Valley with field offices in Washington D.C., New York, London, Tokyo, Utrecht and Frankfurt.