MarkLogic®

# What's New In MarkLogic 9

WHITE PAPER • MAY 2017

The best database in the world for data integration is now even better with MarkLogic 9, our most ambitious release yet. MarkLogic 9 includes major new features to improve data integration, security, and manageability—all designed to make integrating data from silos easier and faster, and to ensure that your integrated data is well-governed and managed.

# Contents

**Introduction** ..... 2

**Faster, Easier Data Integration** ..... 3

Entity Services: Model-driven data integration around real-world entities ..... 3

MarkLogic Optic API: Joins and aggregates over documents ..... 4

Template Driven Extraction: View your documents with lenses ..... 5

ODBC Driver: Improved integration with BI and analytics tools ..... 5

Data Movement SDK: Bulk data movement with Java ..... 5

**Industry-Leading Data Security** ..... 6

Encryption: Advanced encryption at rest ..... 6

Element-Level Security: Granular security at the sub-document level ..... 7

Redaction: Export controls to avoid sharing sensitive data ..... 7

**Improved Manageability** ..... 8

Ops Director: One clear picture of your MarkLogic system ..... 8

**Even More Enhancements** ..... 9

**Getting Started With MarkLogic 9** ..... 10

# INTRODUCTION

MarkLogic® 9 is the most ambitious release to date of the world's best database for integrating data from silos. This is true in the literal sense: MarkLogic 9 includes more new features and updates than any previous release. The new and enhanced features are focused on three core areas: data integration, security, and manageability. MarkLogic was already the best database for data integration, the most secure NoSQL database, and had a robust set of admin tools—MarkLogic 9 elevates these capabilities to a whole new level. These improvements position MarkLogic as the leading new-generation database for integrating all of your data from silos and running modern, transactional applications at scale.

## FASTER AND EASIER DATA INTEGRATION BUILT ON MARKLOGIC'S MULTI-MODEL CAPABILITIES

Traditional relational approaches compromise flexibility and governance by limiting you to a rigid, tabular data model when integrating data. MarkLogic 9 introduces a suite of new features that frees you from those constraints by making data integration easier and more iterative so that you can get more value out of your data, faster. You can bring all your siloed data – whether relational or non-relational – into a single unified platform that is designed for versatility. And, with MarkLogic 9, you are empowered to expose your data however you want. You can view it as documents, as a graph, or as relational data by leveraging the new SQL capabilities. You can even combine models. With this unprecedented flexibility, you can avoid expensive and brittle ETL and better manage the entities and relationships that your business works with.

## IMPROVED DATA GOVERNANCE AND ADVANCED SECURITY FEATURES

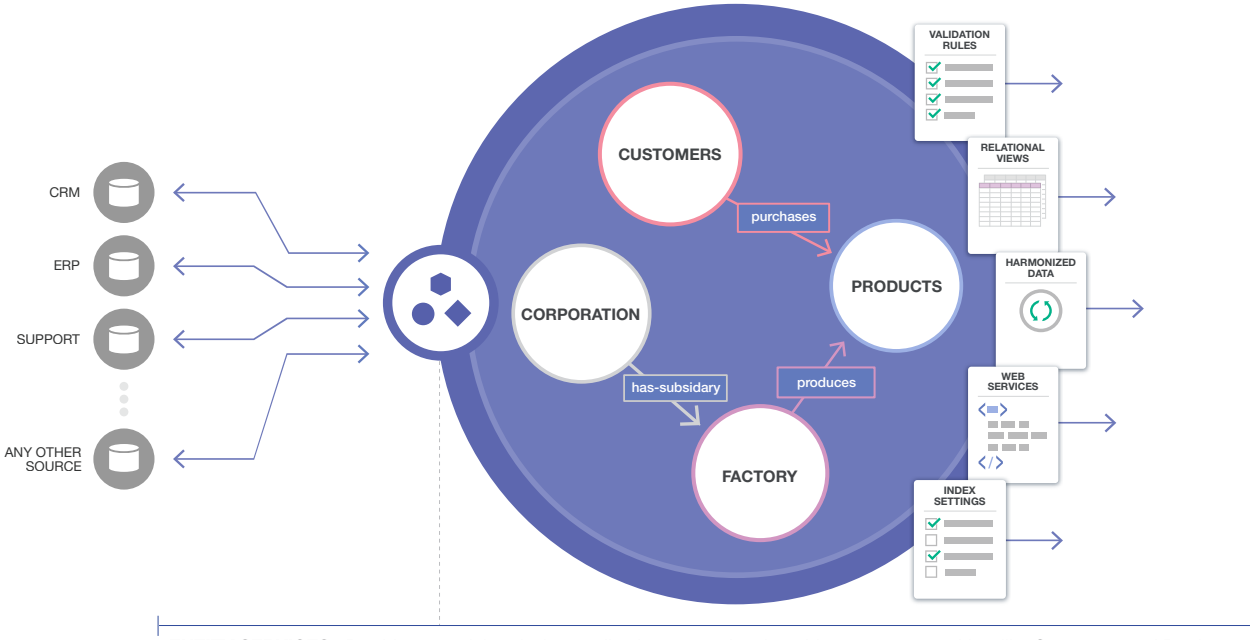
What would be the purpose of integrating data from silos if that data is not properly governed and secured? MarkLogic 9 is designed with new capabilities to govern your integrated data through the entirety of its lifecycle and to secure it against increasing cybersecurity threats. For over a decade, organizations have run MarkLogic in mission-critical environments, and new security features further solidify MarkLogic's position as the most trusted NoSQL database for data governance. MarkLogic 9 is bigger than that though—it is a major step towards making MarkLogic the most secure database, period.

## IMPROVED MANAGEABILITY TO MAKE DBAs MORE EFFICIENT

With the rapid proliferation of applications and services around the globe, DBAs are in need of better tools than ever to keep up with the exponential growth. MarkLogic 9 introduces new features that provide a more unified platform for managing globally distributed MarkLogic database clusters throughout their lifecycles. With a clear picture of the system and improved admin tools, DBAs can increase automation and focus more time on tasks that add value to the business.

## ADDITIONAL ENHANCEMENTS THAT EXTEND MARKLOGIC CAPABILITIES

Beyond the improvements in the thematic areas of data integration, manageability, and security, MarkLogic 9 includes many other additional enhancements. Compliance Archives make it easier to store, manage, and secure historical data using time and event based policies. Tiered Storage, a feature for moving data to less expensive storage, is faster and more flexible. Geospatial gets more advanced and precise. Search gets more robust with expanded language support and other improvements. Query Console gets JavaScript profiling and type-ahead suggestions. And, we now support running MarkLogic on the Microsoft Azure cloud. Not to mention hundreds of other improvements that continue to set the bar high for Enterprise NoSQL.




**ENTITY SERVICES:** Provides a model to design applications around real-world concepts, or entities, like Customers and Products.

## FASTER, EASIER DATA INTEGRATION

Two of MarkLogic's main differentiators are the flexible data model and sophisticated indexes, which together make integrating data from silos faster and less risky. MarkLogic 9 introduces a suite of new capabilities that take advantage of those differentiators to make data integration even faster, easier, and more powerful.

One new feature, **Entity Services**, provides a way to catalog and manage business entities (*Customers* and *Orders*, *Trades* and *Counter Parties*, *Providers* and *Outcomes*, etc.) in as much or little detail as needed, and over time as new data is added. The new **Optic API** provides unprecedented flexibility by giving developers a fluent interface to query all of their data using either JavaScript, XQuery, Java, or Node.js. The Optic API also makes it possible to do efficient joins and aggregates over documents—making MarkLogic the only database in the world that can do this. The underlying technology enabling this capability is **Template Driven Extraction (TDE)**, which makes it possible to define a relational lens over your document data so you can query parts of your data using standard SQL or the new Optic API—without changing the underlying data. You can also define a triples lens over your document data and query with SPARQL or the new Optic API.

The SQL engine in MarkLogic 9 has been revamped from the ground-up to provide faster SQL and **ODBC Driver** access to better integrate MarkLogic with your business intelligence, reporting, and analytics tools. And, what would data access be if you were unable to easily get data into and out of MarkLogic? The **Data Movement SDK** – a new Java library for applications – makes it easy to move large amounts of data into, out of, or within a MarkLogic cluster.

### **ENTITY SERVICES:** MODEL-DRIVEN DATA INTEGRATION AROUND REAL-WORLD ENTITIES

With its flexible data model and ability to store multiple schemas simultaneously, MarkLogic is an excellent database for data integration. But, organizations don't just need the ability to store multiple schemas simultaneously. They also need to query against fixed, predictable aspects of the data that represent real-world things, or entities. Examples of entities include *Customers* and *Orders*, *Trades* and *Counter Parties*, or *Providers* and *Outcomes*.

The problem with traditional relational databases is that they have a static data model and can store

only one schema. This makes managing entities extremely challenging. The context and meaning of data is trapped in database queries, application code, outdated application specs, and entity-relationship diagrams (ERDs)—everywhere *except* in the database. This means it is near impossible to make sense of your data:

*What data accurately represents our Customer? What are the defining properties? How is it related to other entities? Which systems can generate customers? How are customers represented to applications? Which customers do not adhere to the business rules?*

Entity Services provides a better way to manage entities and the messy, changing data from which they are derived. You can think of it as a catalog that defines a shared understanding of your entities and relationships that makes data easier to govern and program. Data owners can use Entity Services to build a model that captures governance policies and data rules right in the database, where they belong. Developers can use Entity Services as part of a model-driven workflow to automatically generate data transformations, validation rules, index configuration, and SQL views, all of which reduce errors and help accommodate inevitable change.

## MARKLOGIC OPTIC API:
## JOINS AND AGGREGATES OVER DOCUMENTS

Document databases provide a natural, denormalized representation of data, where documents represent holistic entities free of joins and complexity. The benefit of this data model is that it maintains context and supports development and search, as all information about an entity is conveniently in one place for efficient retrieval and update. But, many questions also rely on the relationships across entities—and that is when you want to do joins and aggregates *across* documents.

The revolutionary Optic API unites the relational world and the NoSQL document world by performing joins and aggregates across documents. MarkLogic is the only database in the world that can do this. One of the enabling features, Template Driven Extraction, creates relational "lenses" over documents by using templates to specify the parts of a document that make up a row in a view. A template does not change your documents in any way, it just changes how the indexes are populated based on values in the documents.

You can then access your data, viewed through your lenses, using either SQL or the API. With the Optic API, you can use full-text document search as a filter, perform row operations to join and aggregate data, and retrieve or construct documents on output. You can even use the Optic API to combine multiple lenses in a single query, including triples.

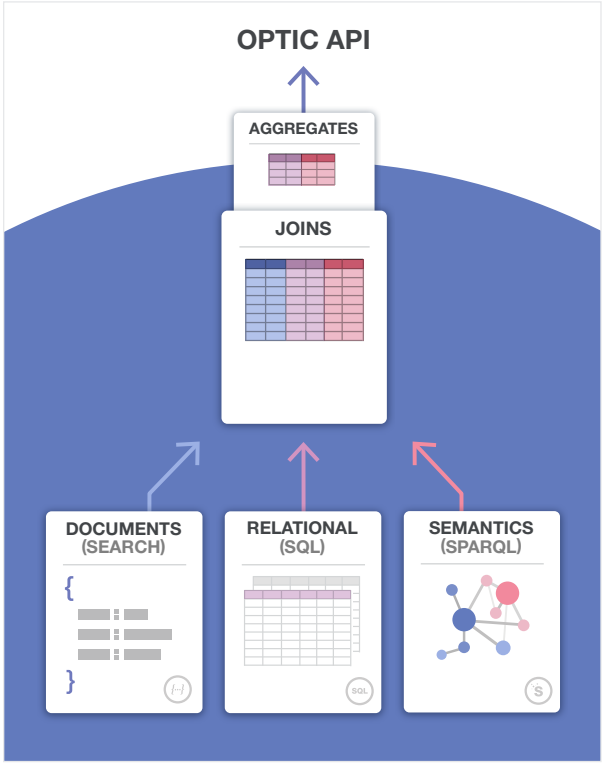



Figure 2: The new Optic API provides a native language query interface to perform joins and aggregates over documents.

Developers benefit from the simplicity of a unified query interface and efficient in-database processing. And, it's available in all of MarkLogic's supported interfaces: Java, JavaScript, and REST on the client and JavaScript and XQuery in the server. Each implementation adopts language-specific patterns so it feels conceptually familiar for developers with relational experience and syntactically natural given existing programming knowledge. For example, there is no need to compose a string and feed it to the appropriate special-purpose query language (SQL or SPARQL). Instead, you build up your query using your chosen programming language.

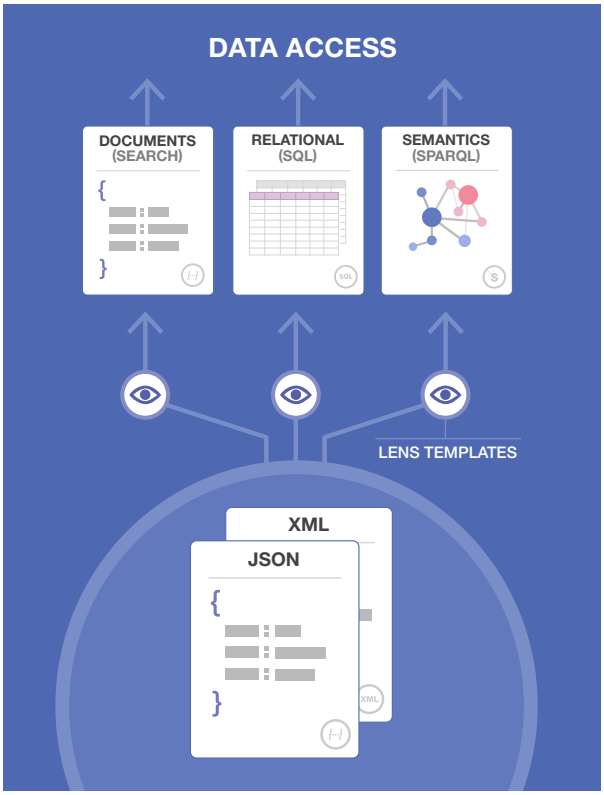



Figure 3: Template Driven Extraction is an enabling feature that allows you to view your documents through relational and semantic lenses.

## **TEMPLATE DRIVEN EXTRACTION:** VIEW YOUR DOCUMENTS WITH LENSES

MarkLogic supports the document model, a flexible, schema-agnostic approach that natively handles rich data including JSON and XML. But, you may want to look at some parts of your data using a relational view, and query that data using the well-known SQL language.

With Template Driven Extraction (TDE), you can overlay parts of your document data with a relational view, without changing the documents themselves. We call this a *relational lens*. Using a Template, you can create a relational lens over documents, specifying which parts of documents make up columns in a view. Inserting that Template into the database creates a SQL view and populates the indexes, so you can query the view with SQL (server-side SQL or ODBC) or with the Optic API. This is particularly useful when you want to create reports and visualizations using tools that communicate using SQL, or when you want to join entities and perform aggregates across documents.

You can also use TDE to define a semantic lens, specifying which data from a set of documents make up RDF triples that are then indexed in the triple index, making them queryable with SPARQL or the Optic API.[1]

With TDE, templates are entirely independent of the documents—they do not change your documents in any way. They just change how MarkLogic's indexes are populated based on values in the documents.

## **ODBC DRIVER:** IMPROVED INTEGRATION WITH BI AND ANALYTICS TOOLS

MarkLogic 9 enhances its SQL engine to provide users a more robust and improved experience using their business intelligence (BI) and reporting tools. The new SQL engine is simpler to set up, scales to more columns (without relying on memory-mapped indexes), and runs queries faster. The new SQL engine drives improvements for the BI investments that many MarkLogic users have made, as they have often standardized around an ODBC connection. MarkLogic has supported ODBC connectivity since MarkLogic 6. Now, by leveraging the revamped SQL engine, the ODBC driver takes advantage of functional and performance improvements within the MarkLogic architecture.

## **DATA MOVEMENT SDK:** BULK DATA MOVEMENT WITH JAVA

MarkLogic makes it easy move data into, out of, or within a database. Customers are doing this at scale to load nightly batches from legacy systems, generate sophisticated reports, or bulk transform millions of documents at a time.

MarkLogic 9 introduces a new software development kit (SDK), as part of its existing Java API. This Data Movement SDK allows developers to easily include bulk data movement in their Java applications. Under the covers, the Data Movement SDK handles all of the intricate orchestration required to parallelize database access and maintain availability, even as the MarkLogic cluster changes (like when adding nodes).

1 Download the Flexible Data Model Datasheet and the Semantics Datasheet for a better understanding of MarkLogic's multi-model capabilities.

The Data Movement SDK complements the existing Java Client API with an asynchronous interface for reading, writing, and transforming data in a MarkLogic cluster. It enables integrations with existing ETL-style workflows, such as when writing streams of data from a message queue or to transfer relational data via JDBC or ORM (object-relational mapping).

Some of the key Data Movement SDK use cases include:

- Ingesting data in bulk or streams from any Java I/O source, leveraging your entire cluster for scale-out performance
- Bulk processing documents in situ in the database using JavaScript or XQuery code invoked remotely from Java and executed within the database, close to the data
- Exporting documents based on a query, optionally aggregating to a single artifact, such as CSV rows

## INDUSTRY-LEADING DATA SECURITY

Data security and data privacy are top-of-mind in every enterprise and MarkLogic 9 has significant new capabilities to support both. **Encryption** allows data, configuration information, and logs to be fully encrypted at rest. And, security is more granular with **Element-Level Security**. This provides access control at the level of XML *elements* or JSON *properties* within documents, in addition to the existing document-level security. Furthermore, **Redaction** addresses privacy concerns by making it possible to remove or mask information when importing, exporting, or copying data. This prevents leakage of sensitive information to unauthorized users.

### **ENCRYPTION:** ADVANCED ENCRYPTION AT REST

Rising internal and external security threats, expanding compliance requirements, and cloud computing are just a few reasons why encryption is critical. For example, imagine the common scenario of a system administrator that has full access to files on a server, the unpredictable threat of a hacker that obtains access, or a database hosted by a cloud provider outside your control. Without encryption, or even with file system encryption, the system administrator, cloud operator, or hacker could access or modify files—including the files that comprise your database.

Encryption is a new feature in MarkLogic 9 that allows data, configuration, and logs to be encrypted transparently. This means that MarkLogic files *on disk* are encrypted at rest, preventing access from outside MarkLogic. This feature requires no modification to applications developed on MarkLogic databases. Encryption in MarkLogic is comprehensive and flexible, allowing the administrator to choose which databases to encrypt, if configuration should be encrypted, and if logs should also be encrypted.

Encryption works with keys generated by MarkLogic or an external Key Management System (KMS). Using an external KMS offers an additional layer of security by separating controls between the MarkLogic administrator and the security administrator who controls the KMS. Regardless, by encrypting logs and configuration, no unauthorized users, not even MarkLogic administrators, can delete or modify auditing events in log files, thus preventing them from erasing their traces of wrong doing.

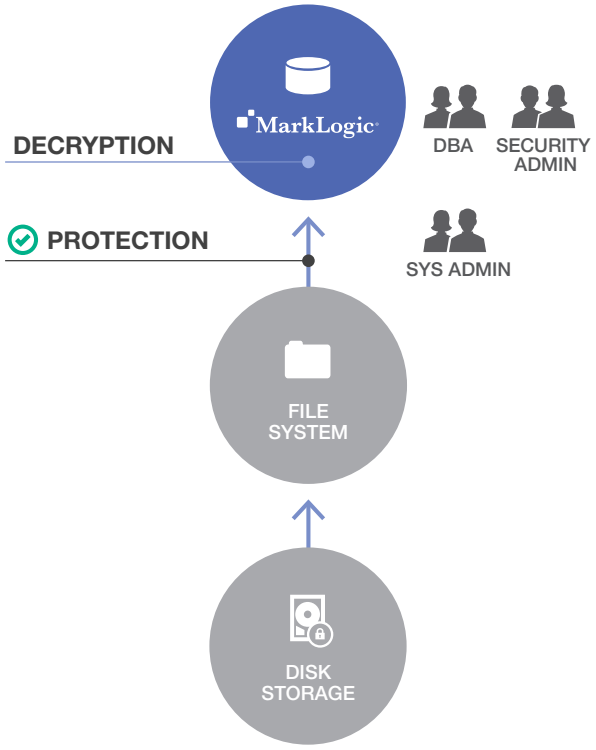



Figure 4: Advanced Encryption prevents access from outside MarkLogic by ensuring files resting on disk are fully encrypted.

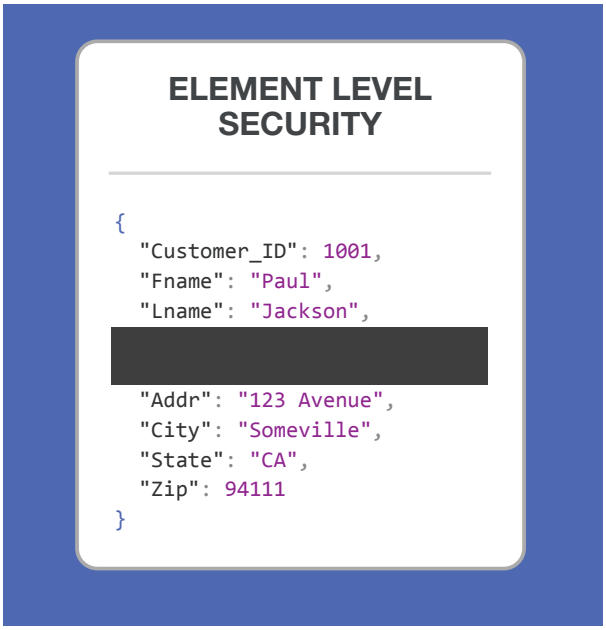



Figure 5: Element-Level Security provides real-time control of who sees what inside documents.

## **ELEMENT-LEVEL SECURITY:** GRANULAR SECURITY AT THE SUB-DOCUMENT LEVEL

Many databases have all-or-none data access, which is not sufficient to protect against today's cybersecurity threats. Today, organizations need to define who can see what data at the most granular level. For example, imagine an application that exposes customer credit card information and personally identifiable information to a call center operator. Exposure of that information, even to valid application users, may violate privacy regulations and put your organization's data at unnecessary risk. Controlling data access at the database-level eliminates these risks and prevents exploitation of application or network flaws that may disclose sensitive data. The security controls are at the level of the database, and not something that application developers need to build into their code at the top of the stack.

MarkLogic has always provided role-based access control at the individual document level. MarkLogic 9 goes further by allowing security administrators to apply these same controls to individual parts of a document. Element-Level Security feature provides access control at the level of JSON properties or XML elements within documents. This means that specific information inside a document may be hidden from a particular user based on the user's role, while still providing access to other information in the document.

Element-Level Security is very flexible in the way it describes what elements are to be protected. For example, an administrator can protect a social security number wherever it happens to appear within the structure of a document (i.e. regardless of schema). The administrator does that using a rich, industry-standard, path expression rather than a rigid specification.

These expressions can include attributes of an element, not just the element's name. For example, an element may be marked with a classification attribute of secret. A more restrictive security rule can be applied to any elements that have that attribute.

## **REDACTION:** EXPORT CONTROLS TO AVOID SHARING SENSITIVE DATA

Sometimes, cyberattacks are carefully thought out and executed. Other times, data is put at risk by accident. For example, imagine the common scenario of testing applications using sampled data from production. A developer or test engineer may have access to real credit card information or personally identifiable information. Today, even that accidental exposure of that information may violate privacy regulations and put your organization at risk.

Redaction eliminates the exposure of sensitive information by removing existing information or replacing it with other values in order to prevent leakage of information that is not required for end-users to execute their duties.

Redaction uses rules to define what information to remove when exporting data. The process is simple, flexible and secure. First, a MarkLogic security admin creates redaction policies that contain rules defining which sensitive information should be redacted and how. Then, the admin chooses which policy to apply when running an export. This process allows admins to generate data sets for different purposes, such as development or data analysis, that have different treatment of sensitive data. For example, a dataset aimed at developers may get a random credit card number while a dataset used by analysts will not have credit card information at all. In addition, the rules and actions are logged, ensuring that you can later audit all export activity.

Figure 5: Redaction is similar to element-level security, but is designed to control who sees what data when the data is exported from MarkLogic.

## IMPROVED MANAGEABILITY

MarkLogic 9 introduces a number of new features that make MarkLogic easier and more efficient to manage. **Ops Director** is a single pane of glass to visualize and manage multiple clusters and multiple product versions, across environments, and across different user groups. **Telemetry** makes support more automated by sending diagnostic and system-level information about your MarkLogic cluster to the MarkLogic support team, reducing case resolution time and manual overhead.

### **OPS DIRECTOR:** ONE CLEAR PICTURE OF YOUR MARKLOGIC SYSTEM[2]

While infrastructure costs are falling, personnel costs are rising, and today most organizations have DBAs managing more and more data and servers than ever before. It is critical to make DBAs more efficient at their jobs and modern databases must be as easy to manage as possible. For that reason, MarkLogic 9 has a fully redesigned experience for administrating clusters.

With MarkLogic 9, Ops Director provides a single pane of glass for MarkLogic administration to make it easier to manage MarkLogic, for both experienced MarkLogic DBAs and those just learning. Ops Director presents a consolidated view for DBAs via rich visuals and dashboards. It streamlines monitoring and troubleshooting of clusters with improved alerting, log searching and performance reporting. And, you still get the benefit of MarkLogic's enterprise-grade security with robust Role-Based Access Control and information security. With Ops Director, it is easy to detect problems before they occur, bringing problems to attention without noise and provides learning and analysis opportunities with centralized data collection, delivery and storage.

Here is a short list of some of the things you can do with Ops Director:

- View a dashboard of events happening in your cluster
- Get real-time alerts on key events
- Monitor clusters, hosts, databases and app servers
- View and respond to error logs
- Analyze problematic hosts
- Communicate with MarkLogic Support
- View a management dashboard showing resource utilization
- Analyze performance of disk I/O, CPU, memory, network, databases, and servers
- Track current task statuses for common maintenance tasks
- Manage console settings for security and licensing, connectivity, and telemetry

### **TELEMETRY:** BETTER, FASTER SUPPORT

Telemetry is part of our continuous effort to provide better and more timely support by automating the collection of diagnostic data about your MarkLogic clusters. MarkLogic has a strong track record of providing best-in-class customer support, and Telemetry continues to help by further reducing the turn-around time.

2 Note: The initial release of Ops Director will follow the 9.0-1 server release.

Telemetry is an opt-in feature. When it is enabled, it collects, encrypts, and sends diagnostic and system-level information about a MarkLogic cluster to a secure MarkLogic destination. For privacy and security, user data, application logs, and access logs are never sent. You have complete control to decide what data is sent: Log data, metering data, configuration files, and support requests. And, you can choose the data granularity to send in order to accommodate environments with low or spotty bandwidth.

## EVEN MORE ENHANCEMENTS

Beyond the improvements in the core areas of data integration, manageability, and security, MarkLogic 9 is full of additional enhancements that you will appreciate every day. Geospatial gets more advanced with increased precision. Search gets more robust. Tiered Storage and Bitemporal become easier to implement for compliance. Beyond that, there are hundreds more improvements that continue to set the bar high for Enterprise NoSQL.

- **Compliance Archive** NEW – Protect temporal documents against deletion, updates, and wipes using time-based or event-based policies, and to save documents to WORM storage. This feature works with Encryption to ensure that data cannot be changed on disk by a system administrator.
- **Azure Support** NEW – An Azure image template and recommendations for your deployments on Microsoft's leading cloud platform.
- **Geospatial** – More powerful and precise with the addition of Geospatial Region Search and Double Precision to support applications with rigorous geospatial requirements.
- **Search** – A new plugin API for tokenization and stemming that lets you use any external tokenizer or stemmer for any language, so you can build better search across languages. MarkLogic 9 also adds enhancements to *near* search and *wildcard* searches.
- **Enhanced Tiered Storage** – More flexibility and better performance for Tiered Storage. Taking advantage of MarkLogic's indexes and load balancing, documents can be tiered based on rules (queries). And, queries across tiers are more efficient due to query partitioning.
- **JavaScript** – Server-side JavaScript has been updated to the latest version of JavaScript, ES2015, which provides new syntax and standard libraries to make code easier to write, read, and debug.
- **Query Console** – Type-ahead suggestions automatically suggest functions, namespaces, in-scope variables, and keywords, depending on the current context. Also, a new Profile mode to better understand query execution, generating a report about what the input code is doing under the covers, including where time is spent.
- **Node.js** – Updated to track new MarkLogic 9 security and geospatial features.

## GETTING STARTED WITH MARKLOGIC 9

MarkLogic 9 continues to set the bar high as the only Enterprise NoSQL database, and now it is even more capable to help you integrate your data from silos.

- **Download MarkLogic 9** from the MarkLogic Developer site
- **Read the release notes** in the documentation for more details about MarkLogic 9
- **Take a MarkLogic University course** covering all of what's new or just on the new data integration features

MarkLogic 9 is a free upgrade for all customers with an active support contract. For information on how to upgrade from previous releases, please review the release notes. The documentation also provides instructions for running MarkLogic 9 in any environment.

If you need additional support to upgrade to MarkLogic 9, MarkLogic Consulting offers an **Upgrade Accelerator** that is designed to help existing customers put the powerful features of MarkLogic 9 into action. To get started with your MarkLogic 9 Upgrade Accelerator, contact your MarkLogic Account Representative or Consulting Director. Alternatively, please give us a call at 1-877-992-8885 or email us at consulting@marklogic.com.

For new customers interesting in purchasing, please visit marklogic.com or send an email to sales@marklogic.com.

**© 2017 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.** This technology is protected by U.S. Patent No. 7,127,469B2, U.S. Patent No. 7,171,404B2, U.S. Patent No. 7,756,858 B2, and U.S. Patent No 7,962,474 B2. MarkLogic is a trademark or registered trademark of MarkLogic Corporation in the United States and/or other countries. All other trademarks mentioned are the property of their respective owners.

MARKLOGIC CORPORATION
999 Skyway Road, Suite 200 San Carlos, CA 94070
+1 650 655 2300 | +1 877 992 8885 | www.marklogic.com | sales@marklogic.com

999 Skyway Road, Suite 200 San Carlos, CA 94070
+1 650 655 2300 | +1 877 992 8885
www.marklogic.com | sales@marklogic.com