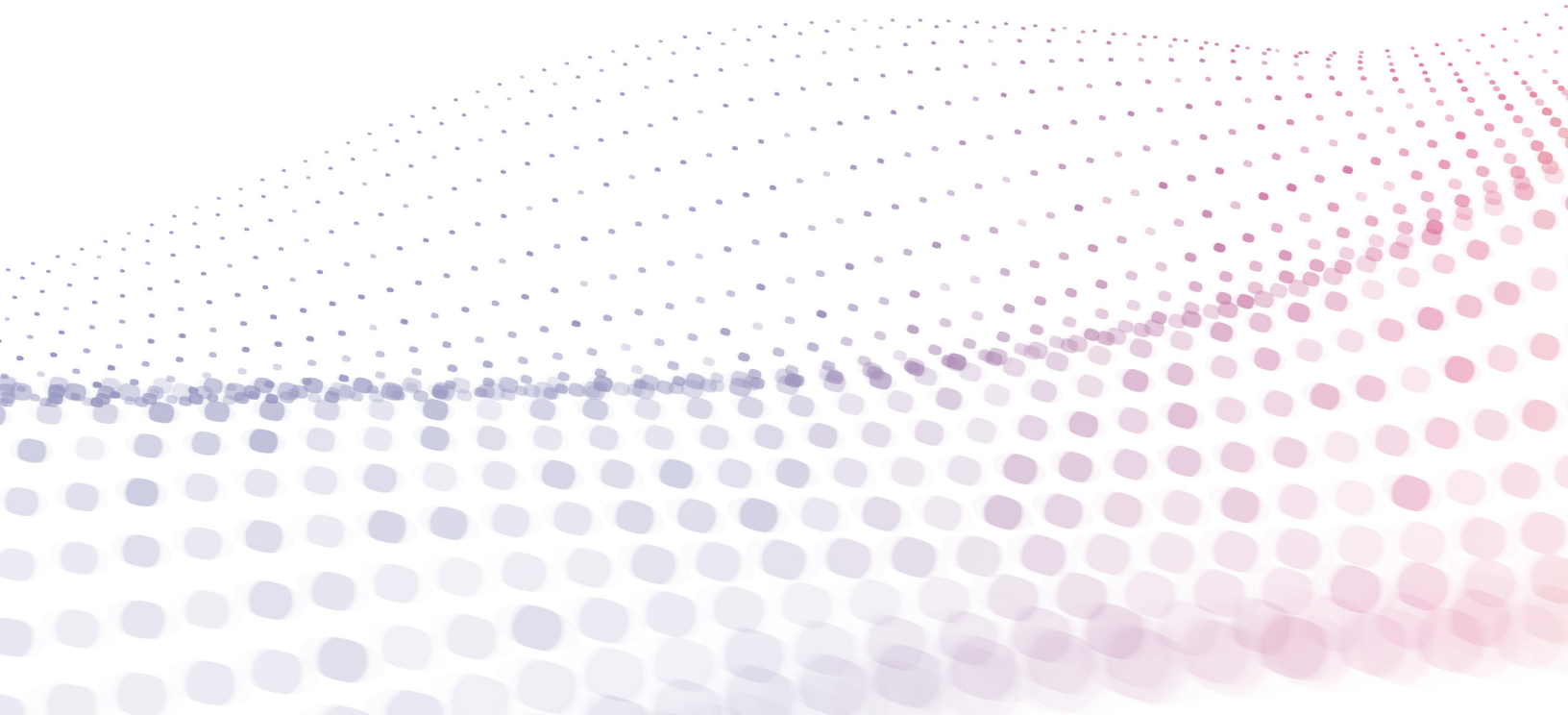
MarkLogic®

Smart Metadata Management

MARKLOGIC WHITE PAPER · JULY 2018

Data has enormous value, but that value that can only be unlocked if the data is truly understood. Understanding data, though, is only possible through metadata – and metadata is usually locked away in silos and its full value never realized. This white paper describes how organizations can unlock the value of metadata, and introduces a better data modeling and architectural approach for smart metadata management using the MarkLogic® database.



Contents

> “ The MarkLogic database enables a smarter approach to metadata management that keeps data and metadata together.”

INTRODUCTION

Data: Businesses require it, run on it, and revolve around it. But, the sad truth is that organizations often focus so intently on bits and pieces of their data that they fail to understand it as a whole. Instead, various collections of data are separated and walled off. If these collections are allowed to interact at all, it is typically done through nightly ETL jobs that are both slow and prone to failure.

Beyond being a headache for DBAs, data silos also result in a failed understanding of the data. Governance and data quality are sacrificed. Audits become things to be feared. The organization fails to thrive.

Enter metadata: If managed properly, metadata enables organizations to integrate their data, govern it, and use it to meet a variety of otherwise intractable business needs. The problem is that metadata and all of its attendant value is as locked away in silos, with no cross-enterprise way of managing it and no way to unlock its full value.

Fortunately, the MarkLogic database enables a smarter approach to metadata management that keeps data and metadata together. In this white paper, we look at what metadata is and why it matters. We then discuss how MarkLogic provides smarter metadata management, going into depth about MarkLogic's multi-model approach, search and indexing, and the Operational Data Hub pattern that makes it easy and fast to bring data and metadata together, govern it, and get value out of it faster than ever.

WHAT IS METADATA?

Metadata, in a very broad sense, is data about data. That is a useful entrée into discussions about metadata, but it barely scrapes the surface.

Part of the challenge with defining metadata is that it means so many things to many people. In our survey of MarkLogic customers, we found that what metadata means is dependent on how an organization creates, manages, and consumes data.

Those who consume data as end users are likely to think of metadata as being ontological or taxonomic in nature, or pertaining to the relationships in the data. Data architects and database programmers, on the other hand, tend to think of metadata as it pertains to data models, the history and maintenance of data artifacts, or workflows within the organization.

Then, there are differences in how metadata is defined depending on the industry or use case. With IoT use cases in financial services, for example, there is metadata about ATMs, payment processing terminals, smartphones and watches used for payment, and also physical security sensors. For most of these components, there is metadata for the purposes of IT asset management, only some of which is provided by the device itself: Firmware version, physical location, model, error logs, serial number, owner, replacement parts provider, operating manual url, etc. Then, there is operational metadata that is indirectly related to the device: Security certificates, schemas, schedules, etc.

The truth is that metadata is many things. The lines between metadata and data are fluid. Depending on the point of view, one organization may call something data and another organization call it metadata. Either way, it is important to identify what can and should be captured, and how to manage it.

WHY DOES METADATA MATTER?

SEARCH & ANALYTICS

Data is the most valuable asset for many organizations. It stores a record of the past – and with proper data analytics, it can provide insights into the present and help chart a course to the future by letting a business focus its efforts on areas that are proven to be successful. Indeed, properly leveraging data is vital to the health of a business, and metadata is key to getting that leverage across every type of data process. Why?

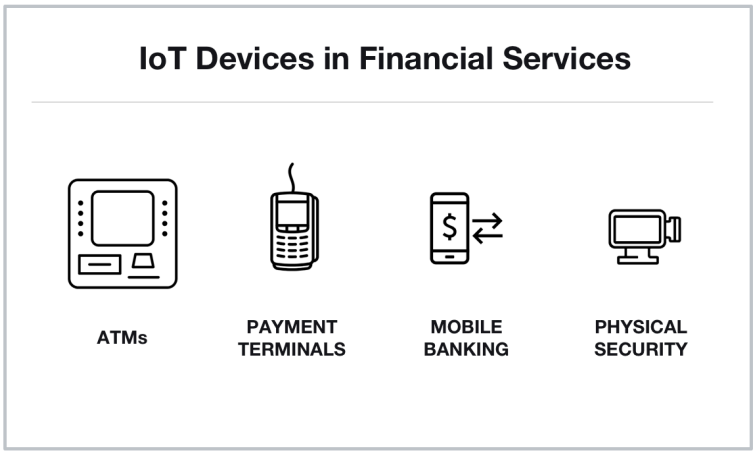


Figure 1: IoT devices in financial services provide a good example of the wide variety of metadata that is available to be leveraged.

Because metadata helps provide a unified, actionable, 360-degree understanding of data. Without an understanding of the data, there is no hope of properly leveraging it – for data analytics or for any other purpose.

GOVERNANCE

Without proper governance, data can quickly get messy, difficult to use, corrupted, or even compromised. Metadata allows for better data governance, and thus drastically increases confidence in the reliability and security of data. We will go into more detail later, but at this point what is important to understand is that MarkLogic has the metadata management tools and features needed to ensure proper data governance, including everything from tracking provenance and lineage to securing data at the most granular level.

Metadata about the provenance and lineage of your data tells the story about where your data comes from, how it changed over time, and who made those changes. That kind of metadata is always important to know, but ETL processes that alter data can lose this kind of administrative metadata. This metadata management problem can be solved by storing data and metadata together, along with the original source data. No matter how the data changed, or how often, the metadata provides the full audit trail.

Proper metadata management is also critical for data security. It helps control who has access to what data. Metadata is used to define users, their roles, the permissions attached to those roles, and how all of that changes over time.

Proper metadata management is also critical for meeting the various regulatory requirements imposed on the owners of data, as is the ability to *prove* compliance with regulatory requirements. If you do not store metadata along with the original data, how can you audit your data integration process, especially as business rules or regulatory requirements change? Most organizations do not have great awareness of where their data is from, how it changed, and how it is being used. And, no one knows what a regulator or auditor will ask in the future (not even them).

WHY DO DATA ARCHITECTS CARE?

Every data architect should care deeply about data integrity – and *metadata* integrity. And yet, creating and managing the appropriate metadata across all of the silos typically found in an enterprise-scale organization is a Herculean task.

Data architects are relied upon to be good stewards of both their data and their business, which means using metadata to provide both efficiency and agility. Just think how much efficiency and agility would be gained if developers knew exactly how to access certain data with a specific API, and how that data is governed. Or, if a compliance director could immediately answer the request of a regulator to see the history of a trade. Or, if the business analyst could immediately see the status of a complicated licensing agreement. Data architects can make all that possible with smart metadata management.

> “If you do not store metadata along with the original data, how can you audit your data integration process, especially as business rules or regulatory requirements change?”

ROADBLOCKS ON THE PATH TO SMART METADATA MANAGEMENT

Enterprise data environments are messy. Data is scattered across multiple silos, making it both difficult and slow to access. Even worse, similar records representing the same real-world entity are frequently present in multiple silos, leading to frequent batch jobs aimed at re-synching the various databases (at best) or outright breaches of data integrity (at worst).

Today, most enterprise data sits in relational database management systems (RDBMS) that are designed for structured data. They use strict schemas that pre-determine that structure and they are difficult to change. This inflexibility leads data architects to cheat on the schema rather than spend time trying to improve it. Data will be stuffed into existing columns just to make things work, or columns and tables will be haphazardly added. Or, the data (and metadata) is just thrown out during ETL jobs because it does not fit that pre-determined schema. The inflexibility of relational databases causes data quality problems and prevents organizations from getting the most value from their data.

None of which is to blame the victim, so to speak. The fault lies with the relational paradigm, not with developers or data architects. When the only tool you have is a SQL hammer, every bit of data (and metadata!) looks like a nail. What developers and architects really need to get a complete picture of their data is simple—a better strategy to manage metadata.

SMART METADATA MANAGEMENT WITH MARKLOGIC

MarkLogic is an Enterprise NoSQL database designed to integrate data silos better, faster, and with less cost. MarkLogic is a great database for integrating data and metadata because it is easy to load data in with a multi-model approach, fast to access that data with industry standard APIs and query languages, and it has enterprise-grade security.

Those three factors – the multi-model approach, search and indexing capabilities, and data security – all combine to make MarkLogic an outstanding choice for metadata management.

THE MULTI-MODEL APPROACH

There are two complementary ways of understanding what it means to be a “multi-model” database.

First, a multi-model database makes it possible to handle multiple different models for similar business entities. For example, imagine that different business units within an organization all have their own schema – their own model – for describing a User record. A multi-model database is designed to handle each of the different models.

Second, a multi-model database offers multiple modeling techniques to represent data, while still having a single integrated back-end. In MarkLogic’s case, it is documents and semantic triples. Both the document model and the triples model are powerful enablers of metadata management because it means the metadata can be stored right alongside the data itself.

To show the power of documents and triples, let us look at a quick example that compares how a row change history field is stored in a traditional RDBMS versus a multi-model database.

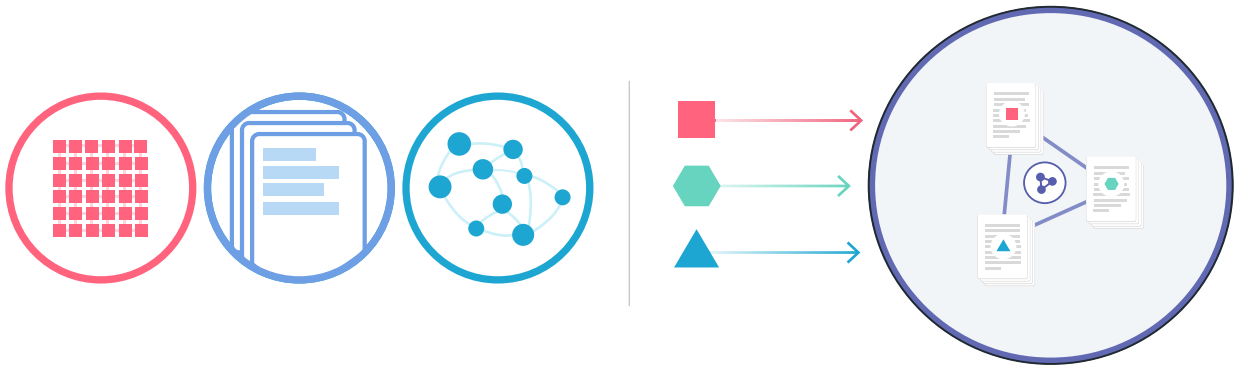


Figure 2: The MarkLogic database uses a multi-model approach to integrate all of your data and metadata together in a single, unified platform.

Suppose you want to know, for any individual record in your database, when (if ever) it was updated. To do this in a relational model, you must add a **last_updated** column. And, because relational data is stored in tables, if you have ten million rows you would add ten million new 'cells' to your data. If only a few records will be updated, that is a lot of null cells (or sparse data, if you like) and potentially a lot of indexing and statistics to update. It is a lot of work when it should have been a simple task.

Now contrast this to a multi-model database like MarkLogic. As a multi-model database, MarkLogic stores data as documents, and any branch in any document can have new branches added to it any time. It has complete flexibility about how documents are extended – or aren't. In our example, if only ten records had been updated, only those ten documents get the **last_updated** value. The other 9,999,990 records are not saddled with an empty field.

SEARCH AND INDEXING CAPABILITIES

The second factor in MarkLogic's superior metadata management is its search and indexing capabilities. We will talk more about MarkLogic's Universal Indexing later on. For now, it is enough to know that when data is loaded into MarkLogic, an index is created that covers all data and metadata.

How do better indexing and search capabilities improve metadata management? It is all about speed and efficiency.

One of the common problems with metadata management across a highly-siloed business is that it makes it difficult to access. If someone wants to gather a complete picture of an entity such as a customer, it is impossible because the information is not all in one place. And, even if it is, it is not searchable. This problem does not exist with MarkLogic because all the data is indexed and searchable in the same way that you search Google. The ability to quickly look at the data and metadata together aids the whole process of data modeling and governance.

MarkLogic's Universal Index is also extremely fast. It is built into the core of the MarkLogic database, and is designed for fast query execution that rivals other leading databases. This is particularly important with metadata management efforts that involve large data sets.

DATA SECURITY

Since data is a high-value asset, it is important to protect it from unauthorized access. Obviously, data must be protected from *external* intrusion. But, it is equally important to protect it from unauthorized *internal* access.

The MarkLogic metadata management approach includes proper security controls. This is because Role Based Access Control (RBAC) is easily implemented and, in fact, becomes a part of the metadata itself. RBAC shields sensitive data by ensuring that the right people see the right data at the right time, and do not have access to sensitive information to which they're not authorized.

This RBAC security is ideal for the organization-wide metadata management that MarkLogic provides. Because MarkLogic stores metadata with its original data, it is easy to provide appropriate, granular access at the document and even sub-document level (similar to "cell" level security in a relational database).

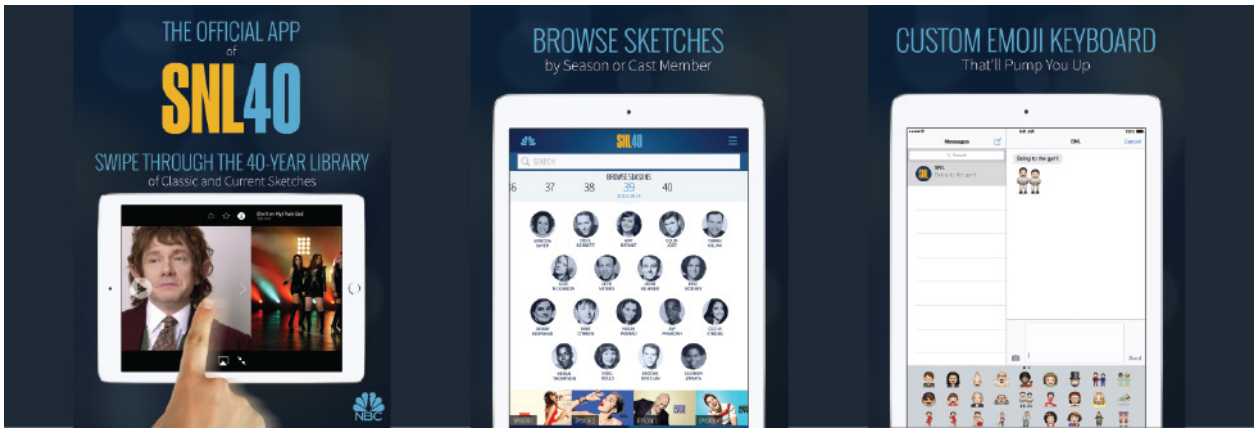


Figure 3: The SNL 40 app is based on smart content generated with the MarkLogic database, letting fans quickly and easily find favorite videos from the last 40 years. The app includes a recommendation engine, driven by MarkLogic Semantics, that adapts to fans' preferences to help them discover new characters and sketches they weren't aware of previously.

MARKLOGIC METADATA SUCCESS STORIES

MarkLogic powers better metadata management for organizations across a wide variety of industry verticals. Capturing some kinds of metadata, such as when a record was created, for example, is universally important to know regardless of vertical. But, each industry does have its own unique kinds of metadata and use cases, as we already mentioned with the financial services IoT example.

In the life sciences vertical, companies are interested in the kinds of real-world evidence that metadata can represent – things like pathology, and how data was collected, and from whom. Using an example from the entertainment industry, movie studios want to know who directed a movie, and how long it runs.

Another example from the entertainment industry provides a quick case study on MarkLogic. The Saturday Night Live 40th Anniversary app built by NBC Universal runs on MarkLogic, and is an example of one way in which smart metadata management helps deliver business value. While the main content of the app is the large binary video files, what makes the app unique is the metadata about that content. The app makes extensive use of a wide variety of metadata about each skit – the cast members, the guest, the air date, and more – and how all of that metadata interrelates to create a graph that can be searched and navigated. To deliver more value, the app has a recommendation engine that leverages this rich tapestry of metadata to suggest videos the viewer may be interested in watching.

SMARTER METADATA MANAGEMENT THROUGH THE INTEGRATION LIFECYCLE

To achieve smarter metadata management, organizations need the right architecture. A common architectural pattern that leverages the power of MarkLogic's multi-model approach is the Operational Data Hub (ODH). The architecture for a MarkLogic ODH is built to handle the end-to-end process for managing data and metadata. The architecture supports every step of the process: (1) Data ingestion, (2) Data curation, (3) Applying security controls, and (4) Providing access to the data.

The ODH pattern makes it easy for architects to handle both observe-the-business (analytics, business analyst-focused) and run-the-business (transactional, customer-facing) use cases.

Unlike data warehouses and data marts, which contain limited and sometimes stale data (and still require ETL jobs), ODH implementations can manage data and metadata in real-time. And, unlike data lakes, which ingest raw data but do not curate it, the ODH pattern both ingests and curates data and metadata, making it ready for true operational use. Not only that, but the ODH also makes it very easy to access the curated data and metadata using industry standard APIs. MarkLogic, being a multi-model database, makes it possible to view the data as documents, graphs, or as structured relational data.

With an overview of MarkLogic and the ODH pattern, let us jump into how it can be used as a platform for smart

> “ Aside from enormous savings in time and cost, MarkLogic’s ingest *as is* capability and the metadata it both captures and generates is vitally important to data governance."

metadata management by following the data integration steps in more detail.

DATA INGESTION

MarkLogic ingests data *as is* from any source and indexes it as it is loaded. This capability is vitally important to data integration. Data from Oracle, SQL Server, Db2 mainframes, Hadoop, or any other source can all be quickly ingested into MarkLogic, with one set of unified metadata created along with it.

Compare that approach to a typical attempt at data unification involving a more traditional, RDBMS-based approach. Data architects can spend many months – years, even – in a process that involves examining all existing data and schemas, creating a new schema that will accurately reflect the structure of the current data, and writing and executing all of the ETL jobs necessary to copy data into the new database.

Often, there are parallel processes for the metadata, which is handled separately. The problem is compounded even further if the source data suffers a schema change over the course of the ETL, necessitating trips back to the source data and more ETL jobs. Data ingestion with a relational database is a fragile and a fraught process. They are intolerant of error, and one mistake can mean a long list of fixes and rewrites.

Aside from enormous savings in time and cost, MarkLogic’s ingest *as is* capability and the metadata it both captures and generates is vitally important to data governance. The MarkLogic ingestion process is non-destructive and the source data remains unaltered. It is an additive process, which means full auditability.

Metadata is created and stored with the original data, and stays together through the data integration process. This is a huge differentiator when compared to legacy approaches.

So, how does it work?

MarkLogic's capabilities lie in the power of the document model. The document model is incredibly flexible, and allows for data to be brought in *as is*, and then curated and accessed downstream. Both the *as is* data and curated data reside in databases in the same MarkLogic cluster (see Figure 4).

The data harmonization process used during the data curation phase is made possible by the document model. The document model's flexibility makes it easy and fast to iterate on during the harmonization of data from various sources. We also recommend employing a data modeling approach called the **Envelope Pattern** to harmonize data, which we go into more detail on in this section. Essentially, during the ingestion process, the raw data is stored as either as XML or JSON documents. That raw source data is preserved, but is wrapped in an envelope that contains metadata.

The metadata section of the envelope holds whatever administrative, structural, or descriptive metadata is desired. Typically, this includes metadata to help with proper governance, security, or use of the record. For example, it may describe when the record was created, and who has access to it.

This example shows the structure of the envelope pattern when metadata is added as it is ingested into MarkLogic:

```
{
"envelope": {
    "metadata": [],
    "source": {
          "your data": "goes here"
     }
  }
}
```

And, here is an actual example showing what the envelope pattern looks like when filled in with real data:

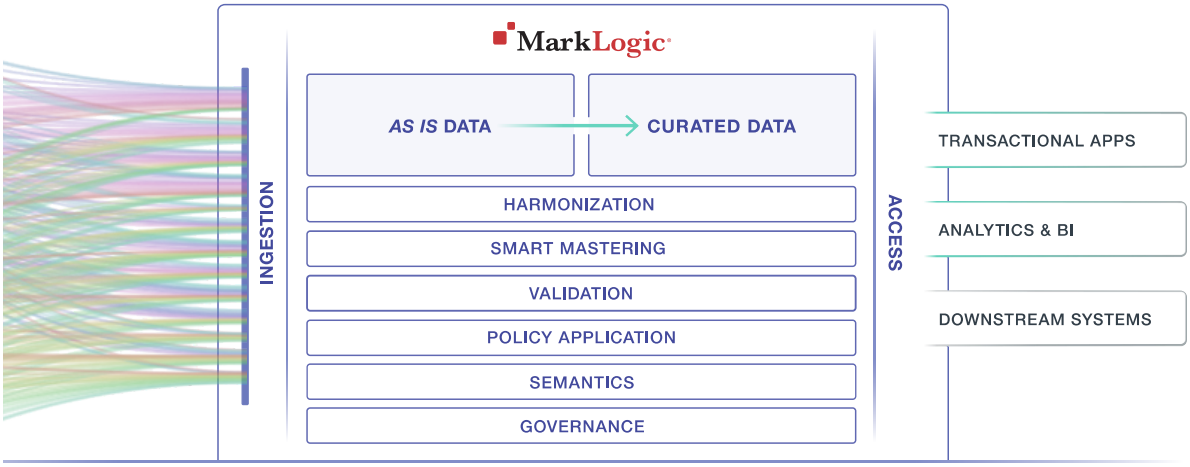


Figure 4: An Operational Data Hub (ODH) is an enterprise architecture pattern that leverages MarkLogic's key capabilities for data operations and analytics. It enables you to load data *as is*, curate that data to form a unified, actionable 360 view, and provide easy ways to access that 360 view.

```
{
"metadata : {
    "Source" : "POS" ,
    "Date" : "2016-04-17" ,
    "Lineage" : "v01 transform" } ,
"source" : {
    "Customer_ID" : 2001 ,
    "Given_Name" : "Karen" ,
    "Family_Name" : "Bender" ,
    "Shipping_Address" : {
       "Street" : "324 Some Road" ,
       "City" : "San Francisco" ,
       "State" : "CA" ,
       "Postal" : "94111" ,
       "Country" : "USA" } ,
    "Billing_Address" : {
       "Street" : "847 Another Ave" ,
       "City" : "San Carlos" ,
       "State" : "CA" ,
       "Postal" : "94070" ,
       "Country" : "USA" } ,
    "Phone" : [
       {  "Type" : "Home" ,
          "Number" : "415-555-6789" } ,
       {  "Type" : "Mobile" ,
          "Number" : "415-555-6789" } ]
}
```

You can see that the envelope pattern has value in and of itself in terms of the metadata that it holds. But, the metadata also drives further value throughout the data integration process.

Keep in mind that the process of data ingestion is iterative. This is important because in an enterprise-scale business, data is often ingested at different times. That is okay. MarkLogic stores ingested data in the *as is* staging database, which allows for multiple ingestion cycles to occur at different times. And, you can do it in batch or real-time using triggers to set off the integration process. These capabilities are important for the next stage of the integration process in which the data is harmonized to provide a 360-degree view.

ANOTHER WAY THE UNIVERSAL INDEX HELPS WITH METADATA MANAGEMENT

MarkLogic's Universal Index indexes all data as it is loaded, including the content and structure of the data. This makes it possible to immediately query the data. This is fantastic for data discovery, and it is also useful for data validation. Data validation is exactly what it sounds like—a process to validate the quality of data that was ingested using a set of rules and criteria. The Universal Index aids this process by allowing data architects to quickly run data validation checks in MarkLogic. They can choose to reject invalid data, or to accept it. If they choose to accept it they can flag it so that it does not get used until it is reviewed, or they can use it for some use cases but not others, or they can make it accessible only for some people and make it invisible to others. As the data validation process happens, metadata is generated and indexed.

CURATING THE DATA

Ingestion, whether through one or many cycles, brings the data into a MarkLogic staging database. By the end of the ingestion cycles, however, multiple records representing the same entity will be present. In order to truly present a universal view of an entity, it is necessary to undergo a process that harmonizes those entity's records. There are other steps that also usually take place, including: Validation, reference data denormalization, deduplication, matching and merging (done using MarkLogic's Smart Mastering feature), and applying of policy. Altogether, this is the process of *curating data*.

MarkLogic's flexible data model and sophisticated indexing, all working together, make data integration simpler with MarkLogic than with traditional, relational paradigms.

The Universal Index, since it has already been created, makes it easy and fast to compare records that constitute an entity. The speed afforded to data architects here is non-trivial, since queries against enterprise-level data sets for data discovery can take several minutes, and sometimes hours, in an RDBMS. When one considers the sheer number of various entities across an organization, the speed of MarkLogic provides a multiplier effect in time savings for database architects in the curation phase.

Because ingested records are quickly accessible, data architects gain an understanding of their data that might not have been practical during pre-ingestion simply due to time constraints. With such an understanding, creating models that accurately reflect all of their data is possible. With that, the creation of metadata provides all of the benefits previously discussed.

The envelope pattern allows for the creation and management of metadata to be stored along with the original data. As previously noted, every document is wrapped in an envelope at the time of ingestion, and at this point the first metadata is created. At this point a document's metadata will primarily be administrative in nature, and contain information such as an ingestion timestamp, data source, data type, and user-specified metadata (such as the standardized first-name label example).

Given that the data architect harmonizing the data has gained a greater understanding of their data through data discovery, now is the time to harmonize specific parts of a document. What data items will be added during harmonization depends on – and are limited only by – the data architect and the business rules they implement.

Further building out our earlier example provides a helpful walk through. It is possible to promote certain properties such as the customer's Zip or the "first-name" property (used in the previous example) in order to create a unified master record for a customer entity. But, we can also use the envelope to store derived or calculated data for our customer, such as the average value of all of that customer's sales transactions. And, although original data is typically stored in the content fields of the record, it may also be useful to parse the customer's bio and update the content to correct any misspellings, for example, and store that original, uncorrected data value for the bio in the harmonization section. This process transforms the data while at the same time preserving it.

When data is promoted for harmonization, it goes into a new section in the envelope:

```
{
  "envelope": {
    "harmonized": [],
    "metadata": [],
    "source": {
          "your data": "goes here"
     }
  }
}
```

Here is an example after promoting certain properties (e.g., "Zip") to the harmonized section:

> “The speed afforded to data architects here is non-trivial, since queries against enterprise-level data sets for data discovery can take several minutes, and sometimes even hours, in an RDBMS."

```
{
"harmonized" : { "Zip" : [ 94111 ,
94070 ] } ,
"metadata : {
    "Source" : "POS" ,
    "Date" : "2016-04-17" ,
    "Lineage" : "v01 transform" } ,
"source" : {
    "Customer_ID" : 2001 ,
    "Given_Name" : "Karen" ,
    "Family_Name" : "Bender" ,
    "Shipping_Address" : {
       "Street" : "324 Some Road" ,
       "City" : "San Francisco" ,
       "State" : "CA" ,
       "Postal" : "94111" ,
       "Country" : "USA" } ,
    "Billing_Address" : {
       "Street" : "847 Another Ave" ,
       "City" : "San Francisco" ,
       "State" : "CA" ,
       "Postal" : "94070" ,
       "Country" : "USA" } ,
    "Phone" : [
       {  "Type" : "Home" ,
          "Number" : "415-555-6789" } ,
       {  "Type" : "Mobile" ,
          "Number" : "415-555-6789" } ]
}
```

Finally, the envelope also contains a section for semantic triples where metadata is stored in the form of RDF triples.

Like the other metadata section, metadata stored as semantic triples can be created at either ingestion-time or during harmonization. If created at ingestion, triples can be derived from join information that is

useful since the document model is a denormalized form of the data. In either case, the semantics of the data allows architects to leverage it for semantic-driven search results and other purposes. The data is richer as a result and can be used for things such as creating advanced semantic search experiences.

When semantic triples are added to the model as another section within the envelope, the structure now looks like this:

```
{
  "envelope": {
      "harmonized": [],
      "metadata": [],
      "triples": [],
      "source": {
         "your data": "goes here"
    }
  }
}
```

Here is an example of the envelope pattern with triples added:

```
{
"harmonized" : { "Zip" : [ 94111 ,
94070 ] } ,
"metadata : {
    "Source" : "POS" ,
    "Date" : "2016-04-17" ,
    "Lineage" : "v01 transform" } ,
"triples" : [
    { "triple" : { "subject" : "Customer
    2001"
    , "predicate" : "placed" , "object"
    : "Order
    8001" } } ,
    { "triple" : { "subject" : "Order
    8001"
    , "predicate" : "contains" ,
    "object" :
    "Product 7001" } } ,
```

```
"source" : {
    "Customer_ID" : 2001 ,
    "Given_Name" : "Karen" ,
    "Family_Name" : "Bender" ,
    "Shipping_Address" : {
       "Street" : "324 Some Road" ,
       "City" : "San Francisco" ,
       "State" : "CA" ,
       "Postal" : "94111" ,
       "Country" : "USA" } ,
    "Billing_Address" : {
       "Street" : "847 Another Ave" ,
       "City" : "San Carlos" ,
       "State" : "CA" ,
       "Postal" : "94070" ,
       "Country" : "USA" } ,
    "Phone" : [
       {  "Type" : "Home" ,
          "Number" : "415-555-6789" } ,
       {  "Type" : "Mobile" ,
          "Number" : "415-555-6789" } ]
}
```

For a practical example of how semantic data is managed in MarkLogic's data integration process, watch the presentation Effective Audit Trail of Data with PROV-O. In that presentation, you will learn how provenance metadata is stored using the Provenance Ontology, or PROV-O. PROV-O is a W3C standard for recording provenance metadata in a machine-readable way. The full coverage of PROV-O is outside of the scope of this white paper, but it is important to note that PROV-O metadata can be captured alongside the data to which it refers by using the envelope pattern.

Ultimately, the envelope pattern provides an excellent way to keep data and metadata together, and leads to a very quick turnaround. This is because faster stand-up times for metadata management systems are possible when metadata is integrated incrementally, as opposed to a multi-year "big bang" project to integrate data.

SECURING THE DATA

Having your data and metadata in one unified place makes it an attractive target. For that reason, MarkLogic provides the fine-grained, certified security that organizations require in order to shield against today's cyber threats. MarkLogic is considered the most secure NoSQL database, and has the same security certifications as the leading relational vendors.

So, how do you put MarkLogic's advanced security features to work? And, more importantly, how can you use metadata to better govern your data?

It is critical to be asking the hard questions at this stage in the process in order to build trust:

- Who has the authority to view it?
- Who has the authority to change it?
- How granular do these permissions need to be?

These questions of trust are overcome by using MarkLogic's approach to metadata management. With the envelope pattern, the metadata is infinitely expandable and fully governable. All records can have metadata in the envelope which indicates both who is allowed to alter the metadata and who did alter the metadata.

MarkLogic does this by using Role Based Access Control (RBAC) to manage visibility and shareability. It creates a trusted environment for metadata management. Enterprise-wide visibility is an essential component of smart metadata management. It gives organizations a universal source of truth that is subject to proper governance, with full data lineage and provenance intact and auditable.

For example, consider records for person entities that have social security numbers. Social security numbers are extremely sensitive and personal, and are also commonly subject to laws and regulations regarding their divulgence. It is important to guard such important metadata using MarkLogic's fine-grained access controls. MarkLogic's anonymization and redaction capabilities can be applied so that the data can be safely shared, whether for the purposes of business intelligence (BI) or for quality assurance and development (QA/Dev).

ACCESSING THE DATA

The end result of ingestion, curating, and securing, of course, is that your data – and all of the metadata about your data – is accessible for both operational and analytics use cases. MarkLogic provides a variety of industry standard APIs out-of-the-box that can be

Figure 5: Example of the same document viewed three different ways. MarkLogic's security controls provide different levels of access to the data. For a BI or QA/Dev export, the data can be anonymized and/or redacted.

leveraged for easy data access. Even for industry and use case-specific applications that require custom connectors, these delivery mechanisms provide an easy foundation to build on.

The MarkLogic metadata management capabilities are highly flexible, and apply to all verticals. But, to further highlight the point that there needs to be a great deal of flexibility in handling metadata, let us look at two MarkLogic case studies. Although in the same industry, these two organizations handle their metadata very differently.

The first company, let us call them Company A, wanted in-depth knowledge of all their movies to provide better viewing experiences and to help with future productions and other marketing efforts. What they wanted was a search application to return relevant movies for any specified content—say, a particular scene showing a car chase in California for example. For Company A, the data comprises all of the movies that they have ever released. The metadata, in this case, is the frame-by-frame content of each movie. Need to find that car chase scene? Company A's search application, using MarkLogic, tells you not only what movie has such a scene, but at what time in the movie it occurs.

Company B, on the other hand, wanted in-depth knowledge of all their movies in order to better manage licensing and re-use of their content. Instead of using MarkLogic to capture metadata about the marked up content of their movies, they capture metadata about file formats, file sizes, running times, and various translations and other versions. This enables Company B to also build a search application using MarkLogic, but the search application is designed to answer very different questions.

The use cases are different but the thing that remains constant is that both applications leverage the flexibility and power of MarkLogic as a platform for smart metadata management.

THE VIRTUOUS CYCLE OF SMART METADATA MANAGEMENT

MarkLogic provides a great platform for storing and searching metadata. This in turn makes it easier to maintain and update that metadata. This in turn makes the platform more valuable. It is a virtuous cycle.

The virtuous cycle is made possible by MarkLogic's flexible data model and indexing. Hundreds of millions of records can be ingested and harmonized far more quickly than would ever be possible across siloed, relational systems. Then, as more data and metadata is added and harmonized, data quality and business results continue to improve.

The overall value of the MarkLogic database improves over time as data volumes grow. This is the opposite of what happens in most data lakes and warehouses, where the value degrades over time as data gets more complex, messy, and siloed. With MarkLogic, you get more value, faster — that's smart metadata management.

KEY RESOURCES

WEB PAGE

The MarkLogic ODH Solution

WEB PAGE

Getting Started with the Data Hub Framework

BLOG POST

Making the Case for Semantic Metadata

ABOUT MARKLOGIC

MarkLogic is the world's best database for integrating data from silos. Read this overview datasheet to learn about what makes MarkLogic different, including some specific customer examples.

Read More

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