Ovum
TMT intelligence | informa

MarkLogic Smart Mastering creates a single view of identity and entities

Publication Date: 18 Jun 2018 | Product code: INT002-000128

Paige Bartley

Ovum view

Summary

MarkLogic remains as consistent as ever in its mission to eliminate enterprise data silos by providing a flexible, multimodel NoSQL database where all data – transactional and operational – can be integrated, accessed, and secured. The company is highly representative of a trend Ovum first identified in 2014; the report *Hadoop, SQL, and NoSQL – No Longer an Either-Or Question* outlined how databases are converging in their capabilities to minimize silos. MarkLogic's Smart Mastering capabilities were announced at the MarkLogic World 2018 conference in May and are designed to help the enterprise resolve the perennial challenge of matching all relevant data to a single identity or entity. As part of MarkLogic 9.0-5 (incremental releases are indicated with a hyphen), Smart Mastering was built into the MarkLogic Data Hub Framework. While Smart Mastering looks to resolve many of the same problems that master data management solutions have always sought to address, its architecture and methodology set it apart from traditional approaches. Through the ability to match all relevant data to a specific identity, Smart Mastering is a critical asset to customer 360-degree initiatives; however, it also has immense implications for GDPR compliance.

Gaining a 360-degree view of customers, data subjects, and entities

Nearly every enterprise organization has a 360-degree initiative, whether that be individual-focused (customer 360-degree) or entity-focused (supplier 360-degree). The holy grail to deliver optimal customer service, more successful cross-sell and upsell efforts, better personalized ads, ideal leverage with partners and suppliers, and maximum customer loyalty has always been to understand every piece of data associated with an individual or entity in a single view. Traditionally, disparate data silos have been the biggest barrier to obtaining this singular view and understanding. MarkLogic, with the ability to ingest any data as is without predefining schemas or conducting ETL, has long sought to eliminate data silos by providing a single NoSQL environment where both operational data and ACID transactional data can be managed in the same horizontally scalable environment. A built-in search engine, with a universal index, has long made it easy to query across data that was previously siloed.

What was missing from MarkLogic out of the box was the ability to master data without using custom code – not just to search and retrieve, but to match and merge all relevant data to a specific individual or identity. Because multiple pieces of data related to an entity may have semantic differences but mean the same thing to a human, it is critical to resolve these differences and create a common view via data harmonization. Once a common view is achieved, data quality issues such as duplicate, incomplete, and conflicting records are exposed. Smart Mastering was built to address exactly this, leveraging fuzzy logic and artificial intelligence to uncover relationships of varying probabilities that were previously undetected.

Human identity, in particular, has long posed a challenge to data mastering. Nicknames, misspellings, and inconsistent nomenclature have often been all but undecipherable to technology prior to the common use of embedded machine learning and sophisticated natural language processing. An entry of "Margaret Smith" in one database may appear with the abstract nickname of "Peggy Smith" in yet another – an association that many humans, let alone a simple keyword search, may miss. Other challenges, such as a John Smith and John Smith Jr. living at the same address with the same phone number, would likely be mistaken as the same identity by simple matching technology. AI, NLP, libraries of multilingual name variations, and weighted scores per field put the "smart" in Smart Mastering and set out to address these issues.

Not your mother's approach to master data management

Traditionally, the approach to achieving this 360-degree view of entities was to implement a dedicated master data management (MDM) solution. Over time, these deployments gained a reputation for long implementation times, high failure rates, difficult-to-calculate ROI, and the tendency to create yet another silo. Traditional methods, based on relational models, had trouble keeping up with the multilayered interrelationships that people have with each other and with things. MarkLogic set out to build mastering capabilities within its unified NoSQL database to address mastering in a novel way.

With the introduction of the MarkLogic Data Hub Framework in the 9.0 MarkLogic release, the company laid the foundation for what would enable Smart Mastering, an in-database approach to data mastering. MarkLogic Data Hub Framework is designed to provide out-of-the-box architecture for building an operational data hub, leveraging MarkLogic as the underlying unified database. The framework is open source and is built around a three-step process:

- ingesting data as is into the database, without ETL, and discovering data
- modeling and harmonizing data directly within the database
- delivering harmonized data to downstream systems and applications, using open APIs

It is step two that allows the overall process of data curation to take place directly within the Data Hub, and that is where Smart Mastering now finds its home. Because harmonization includes normalization, formatting of data for indexing, using semantic triples to enhance understanding, and performing conflict resolution between differing values from multiple systems, this step provides an ideal environment for data to be matched and merged without ever having to physically move data between systems. All relevant data has already been ingested as is into the MarkLogic NoSQL environment, where it can be universally indexed, searched, and curated in a single, scalable repository.

The approach of doing data mastering with an ingest-everything, flexible NoSQL database as the underlying unified repository is unique. What is more unusual is the ability to do data curation directly within that database. Traditional MDM solutions cannot support curation and operational use cases in the same database, and the data movement between repositories needed to curate data is one of the processes that slows down mastering, constraining the volume of data that can ultimately be handled. Because MarkLogic is the only database architecture that Smart Mastering depends on, and it is horizontally scalable, there is really no upper limit to the volume of data that can be mastered.

Smart Mastering also does away with the traditional MDM concept of survivorship, whereby when two or more records are merged, only one is permanently kept. Just because data regarding an individual or entity is no longer accurate does not mean it is no longer of value to the organization. Examples might include changes of address over time or changes in phone numbers. Other cases may exist where it is not clear which piece of data is authoritative, or which piece of data is most useful at a given point in time; for instance, a customer may have several email addresses associated with his or her identity. It is useful to have a current master view but still have access to these older pieces of data. This is the approach Smart Mastering takes, maintaining a master that links back to the data and records that have been superseded. Allowing more transparency and potential downstream leverage of data, this concept also has potential applications in compliance.

Unintended, but positive, consequences for GDPR compliance

Strictly speaking, MarkLogic's Smart Mastering capabilities were not built to be a General Data Protection Regulation compliance solution. However, many aspects of compliance with GDPR – the ability to fulfill critical data subject rights – require technical capabilities that are identical to customer 360-degree initiatives. To comply, organizations need to be able to quickly and accurately identify all personal data in the enterprise data ecosystem associated with an individual data subject. Failure to associate a piece of data with an individual because it, for example, uses a nickname or a former address can run afoul of requirements to erase, rectify, or produce data for the data subject to view. Likewise, for the enterprise to simply apply the required data policies to personal information – lifecycle, access control, or otherwise – it all needs to be found and associated with the right individual. A mastered view of data is required to ensure that no near-duplicate personal data is left behind unidentified, unassociated with (but revealing) the identity of a given individual.

This single view of the individual, via mastered data, is critical to meeting several of GDPR's data subject rights:

- **Article 15 – Right of access by the data subject:** To produce all personal data of a specific data subject, the enterprise needs to be able to correctly associate all relevant data to that singular identity.
- **Article 16 – Right to rectification:** When data subjects invoke their right to correct data, the enterprise needs to be able to update a mastered data copy tied to the entirety of data assets representing those individuals.
- **Article 17 – Right to erasure ("right to be forgotten"):** Perhaps most importantly, when data subjects request right to erasure, all personal data associated with them needs to be erased. If data is not mastered, data copies revealing their identity may escape deletion, lingering in IT systems.

MarkLogic's architectural approach to mastering further bolsters compliance. Elimination of silos makes data easier to find, retrieve, and consistently control at the database level, rather than leaving control functions to higher applications. And by maintaining links to all data copies that are associated with a mastered view, rather than permanently removing them, MarkLogic ensures that there is full transparency into data processing and access; data lineage details, maintained alongside the data, further enhance transparency and clarity into data's use. The traditional concept of MDM survivorship, whereby data copies are deleted once resolved into the master record, is also not wise under GDPR; consistency is favored by the regulation, and data lifecycles should be executed equally for all of a given data subject's information. A full audit trail and records of processing need to be available not just for the mastered view, but also the underlying data elements that went into it.

The Smart Mastering capabilities are just the most recent addition to MarkLogic's existing roster of GDPR compliance qualifications. Minimizing technological silos inherently increases control and transparency of data – core technical tenets of GDPR – but MarkLogic's security functionality is also notable for a NoSQL database offering, as NoSQL has long been perceived as less secure than traditional RDBMS. Element-level (rather than document-level) security and role-based access control, encryption by default, support for external key management systems, redaction, support for external authentication using LDAP or Kerberos, and complete audit capabilities all contribute to making MarkLogic well-suited for GDPR's Article 25 mandate for data protection by design and by default, as well as Article 32's requirements for security of processing.

Appendix

Further reading

2018 Trends to Watch: Data Governance, IT0014-003349 (October 2017)

Hadoop, SQL, and NoSQL – No Longer an Either-Or Question, IT0014-002937 (September 2014)

"MarkLogic becomes NTT Data's preferred NoSQL database," IT0014-003315 (July 2017)

Author

Paige Bartley, Senior Analyst, Data and Enterprise Intelligence

paige.bartley@ovum.com

Ovum Consulting

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Ovum's consulting team may be able to help you. For more information about Ovum's consulting capabilities, please contact us directly at consulting@ovum.com.

Copyright notice and disclaimer

Whilst reasonable efforts have been made to ensure that the information and content of this product was correct as at the date of first publication, neither Informa Telecoms and Media Limited nor any person engaged or employed by Informa Telecoms and Media Limited accepts any liability for any errors, omissions or other inaccuracies. Readers should independently verify any facts and figures as no liability can be accepted in this regard – readers assume full responsibility and risk accordingly for their use of such information and content.

Any views and/or opinions expressed in this product by individual authors or contributors are their personal views and/or opinions and do not necessarily reflect the views and/or opinions of Informa Telecoms and Media Limited.

Ovum

TMT intelligence | informa

CONTACT US

ovum.informa.com

askananalyst@ovum.com

INTERNATIONAL OFFICES

Beijing

Dubai

Hong Kong

Hyderabad

Johannesburg

London

Melbourne

New York

San Francisco

Sao Paulo

Tokyo