MarkLogic®

FINANCIAL REGULATORY REPORTING ACROSS AN EVOLVING SCHEMA

MODELDR & MARKLOGIC - DATA POINT MODELING

MARKLOGIC WHITE PAPER • JUNE 2015 • CHRIS ATKINSON

Contents

Regulatory Satisfaction is Increasingly Difficult to Achieve 3

Introduction to Data Point Modeling 3

MarkLogic Database Platform 5

ModelDR Solution Architecture 5

Integrating MarkLogic With ModelDR 5

Benefits of MarkLogic & ModelDR 6

- Reduce Data Silos
- Improve Data Quality
- Accelerate Information Delivery
- Empower Consumer Information Discovery
- Expedite Regulatory Reporting With Bitemporal Capabilities
- Increase Knowledge Quickly
- Discover Knowledge Across Any Source

Summary 7

REGULATORY SATISFACTION IS INCREASINGLY DIFFICULT TO ACHIEVE

The increasing burden of financial regulation has become the primary reason banks are unable to earn more than their cost of capital. Many of the most systemically important banks are struggling with compliance requirements for improved data aggregation, governance, architecture, and processes. Regulation is directed towards improvement in risk management, reporting, and decision-making practices, but banks are having difficulty in establishing data governance and quality assurance due to data architecture and aggregation challenges.

The frequency of change to financial market data models has made it difficult for financial institutions to achieve harmonious, standardized views of data at a given point in time (often referred to as congruence). The solution for many banks has been to implement extraction and transformation (ETL) processes between multiple silos oriented around distinct business functions. This has diminished data quality and increased regulatory risk. Reconciliation overheads have climbed and the cost of ownership has increased substantially, driven principally by the cost of maintaining these ETL processes.

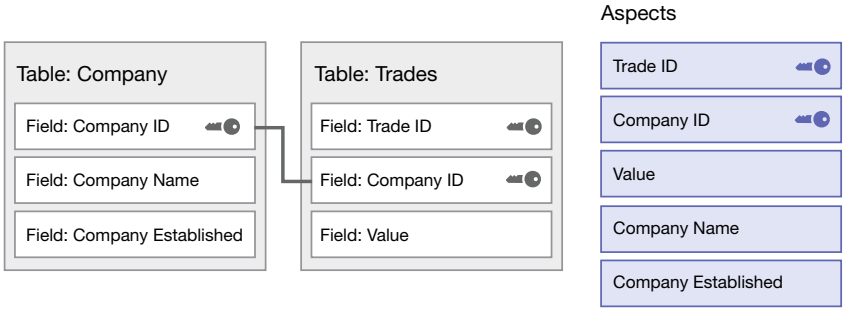
In attempts to reconcile across silos, subsets of information are often collated in relational data warehouses where users query against the schema. Alterations to these schemas require substantial use of operational resources. Schema alterations must be tested, queries validated, and change control processes adhered to. Subsequent re-processing of the data can take hours, days, or even months before information can be produced for the user. Users are often unable to obtain insight from the information as they lack the tools required to visualize complex data environments.

Data congruence becomes more difficult over time as data visibility processes are hard-coded into the ETL process. As the data continues to shift, these processes become increasingly more complex and difficult to maintain.

INTRODUCTION TO DATA POINT MODELING

An approach offering considerable benefits over the more traditional relational data modeling is data point modeling. Supported by XBRL (eXtensible Business Reporting Language), data point modeling is a methodology that decouples your data model from your output format, providing both specific data and flexibility. It can be used to help reengineer relational data structures into a non-relational Data Point Model (DPM). This method involves analysis of the fields within the tables into a set of aspects (facets). When the aspects are created, the relationships between the different aspects are recorded in order to maintain referential integrity (Figure 1).

FIGURE 1



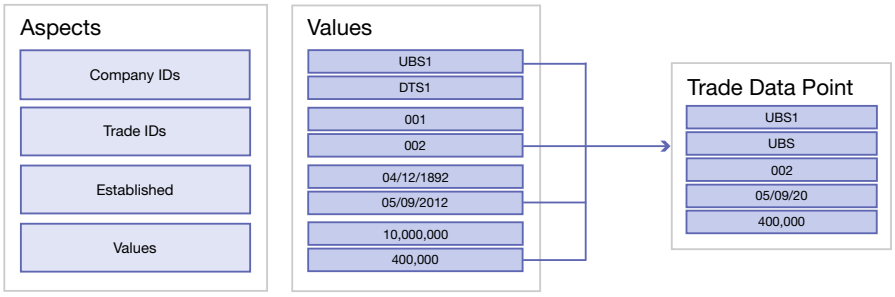


FIGURE 2

A list of all the distinct values is generated against these aspects. These are stored as values. Each value is only recorded once. The data points are then linked to the values within the aspects (Figure 2).

Additional information is also recorded against the aspects when created to help establish congruence across the data from different sources. To better illustrate congruence, consider this example: Data recorded in different locations relating to the same events may be recorded in different forms and within different data models. In order to gain a congruent view of the data, the target data is centralized into a single location. Its model and form are typically standardized through an ETL process into a common data model. Congruence is achieved when the data across all the source systems is proven to be harmonious. That is to say the view at that point in time is complete, information is not missing, and a complete picture is attained.

The key benefits of the DPM approach are as follows:

- Congruence across disparate data sets is achieved quickly without complex and operationally risky transformation processes
- Information converted within the DPM maintains referential integrity
- Users can import a data taxonomy allowing the blending or joining of other data that has been similarly imported into the model
- Rather than querying against a rigidly defined schema, users are instead able to rapidly interrogate the data by using their data taxonomy of choice

Historically the downside to this approach has been that as flexible as the DPM model is, the information is best represented within a non-relational form as Resource Description Framework (RDF) triples and this has created different problems. The generation of the RDF triples from the source data outside the database in effect separates the source data from the derived information within the RDF triples. Subsequent changes to the information sought thus requires re-processing of the sources. What is actually needed is a mechanism whereby the source information can be consumed and RDF triples utilized without necessitating re-processing. Therefore, when reporting demands change, new RDF can be generated and any required reconciliation can be performed in-situ.

Fortunately a database platform and modeling solution now exist which can overcome these historical challenges:

- MarkLogic® – A non-relational, multi-model database platform that integrates both an RDF triple store and document database. MarkLogic stores the original source data in its native form in combination with the generated RDF.
- ModelDR – A data point modeling solution used to visualize and manage the generation and mapping of the RDF and subsequent transformation designs into MarkLogic.

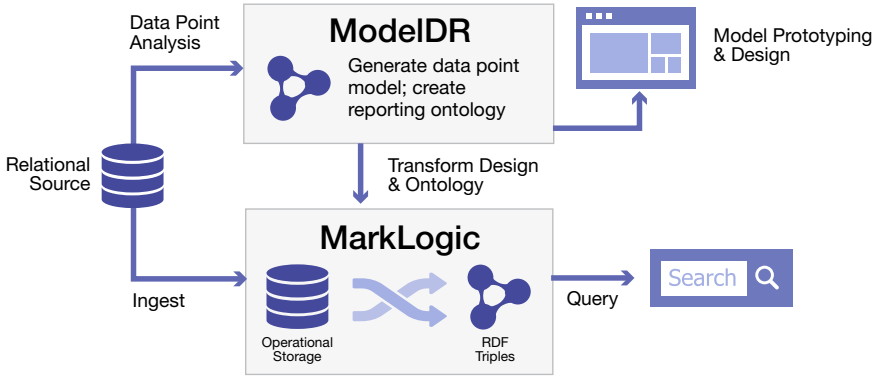


FIGURE 3

MARKLOGIC DATABASE PLATFORM

MarkLogic is an Enterprise NoSQL database platform designed for today's data, including documents, relationships, and metadata. MarkLogic is the only database that can natively store and rapidly query JSON, XML, RDF, and more—providing a single powerful platform for all of your data. The document-centric data model is schema-agnostic, which provides flexibility in modeling data as newly ingested information does not need to first be broken down into table form. With MarkLogic, transactions are stored "as-is," in their natural forms. And, because it avoids data transformation, you can maintain fidelity and context from data conversion while avoiding brittle ETL processes. It makes it much easier to load data from different sources and adapt to changes over time.

MODELDR SOLUTION ARCHITECTURE

ModelDR is a solution that provides users with a graphical interface allowing them to map their relational data sources into a DPM. This DPM is then expressed as an ontology. Next, this ontology is used to generate a transformation map that is used to create RDF triples. This transformation map can be consumed by any database capable of consuming RDF ontologies. ModelDR provides templates based on existing industry ontologies such as FIBO. These templates are combined with the DPM to generate the maps for the ontology and transformation steps. ModelDR users can also use their own business ontologies.

INTEGRATING MARKLOGIC WITH MODELDR

The ModelDR solution encompasses a processing engine with a front end UI capable of analyzing relational sources, importing ontologies, and designing transformations for converting relational data into RDF.

The process for integrating a MarkLogic database platform with ModelDR is as follows (Figure 3):

- **Generate the DPM**
  ModelDR interfacing with MarkLogic will allow users to analyze the relational model and convert it into a DPM. This operation is enhanced with pre-built ModelDR templates for common financial industry ontologies, or existing customer taxonomies may be utilized

- **Create a design layer**
  Data congruence is created in the DPM layer

- **Design transform**
  ModelDR creates an architectural data design for the transform based on SPARQL, the query language for RDF

- **Ingest transform design**
  MarkLogic consumes the transform design to create RDF triples against the source. This can be done retrospectively or plugged into the ingest process itself within the database

- **Query data**
  Using SPARQL queries combined with the generated DPM ontology enables users to query the data

BENEFITS OF MARKLOGIC & MODELDR

MarkLogic provides ModelDR with a flexible and agile database platform, and its schema-on-read approach helps avoid many of the problems associated with aggregating multiple data models into a single operational data store. There are a number of significant benefits to the approach as a whole:

ModelDR provides MarkLogic with a more intuitive and industry focused mechanism for generating and exploring the complex relationships of the reporting ontologies and taxonomies and data. This allows users to manipulate, explore, and generate the required outputs in a reduced time frame and with more accuracy.

REDUCE DATA SILOS

A single common operational data store can be delivered that does not tie users to a single data model. This multi-model capability drastically reduces the degree of effort required when applying source data model changes. The ability to source data natively as-is means that consumers can use MarkLogic as a single aggregated data hub without being forced to alter their data.

IMPROVE DATA QUALITY

Data may be ingested in its native form into MarkLogic avoiding the need for extensive transformation processes. This coupled with data silo reduction significantly reduces the need for "shadow IT" functions. Consumers often individually source feeds, performing their own business transforms and subsequently storing the data within their own silos. These "shadow IT" functions typically arise due to data model restrictions, and loss of agility when attempting to deliver central data hubs. MarkLogic's multi-model capabilities allow consumers to consume a central service without hampering their data model flexibility. The transform designs generated by ModelDR are leveraged at the point of ingest, or can be applied retrospectively to generate the required RDF for subsequent reporting. This is done without changing the original source data. A common data hub is delivered where a single view of the original data and its subsequent derivatives can be sourced and published to with ease.

ACCELERATE INFORMATION DELIVERY

Numerous facets of the MarkLogic database platform expedite the flow of information from creation to consumption. The lack of schema changes allows data to be on-boarded as it changes and the business processes generate information. A lack of rigid relational schemas negates the requirement for vast reprocessing jobs to be conducted. ETL processes are also avoided as the multiple silos previously required to maintain business agility are no longer demanded.

EMPOWER CONSUMER INFORMATION DISCOVERY

The ontology generated by ModelDR is exposed to users, allowing them to easily expose the desired results. Rather than needing to know in advance the facets of the information they wish to expose, users are able to navigate freely through the ontology and expose the information in a discovery-based activity.

EXPEDITE REGULATORY REPORTING WITH BITEMPORAL CAPABILITIES

MarkLogic integrates a Bitemporal query capability into its platform delivering against even the most challenging regulatory demands. Reconciliation overheads are reduced as information from multiple different data models and sources can be queried as they were at the point in time under review. Consumption of existing ModelDR templates or business-generated ontologies expedites the generation of the data views required by customers. Ultimately the generation of the given regulatory reports is performed with significantly less operational overhead.

INCREASE KNOWLEDGE QUICKLY

RDF technologies can be used to create new information within the context of the data. The explicit relationships between data points are used within an inference process to create new explicit relationships, and therefore increase knowledge (Figure 4).

ModelDR provides a lightweight modeling solution that utilizes the data in-place rather than forcing transformation of the data. This delivers a more agile user experience, and faster time to value for data operations. This "in-place modeling" perfectly complements MarkLogic's "store-as-is" capabilities.

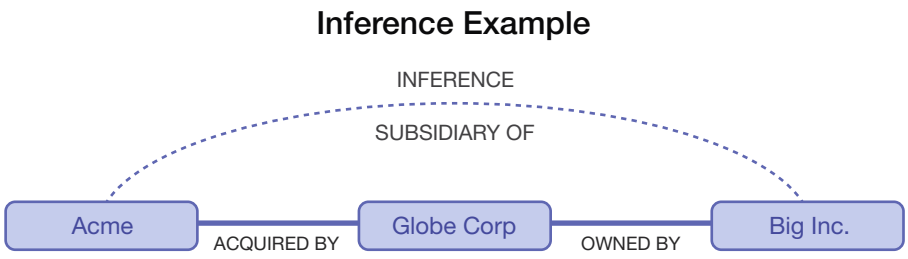


FIGURE 4

Ordinarily, increasing knowledge from data leads to an explosion of the data itself as data is duplicated in numerous positions within the data model. The DPM actively reduces this data proliferation by allowing values to be utilized by multiple data points. De-duplication results are realized as stored values are re-used multiple times.

DISCOVER KNOWLEDGE ACROSS ANY SOURCE

As the structure and metadata are explicit within the data itself, queries can be based on knowledge. New rules and relationships can be simply introduced without requiring structural database changes. These new relationships are then subsequently expressed as RDF within the MarkLogic database layer.

The DPM enables ModelDR to understand any data schema, untangle it and subsequently load solutions into a MarkLogic database platform. MarkLogic enables the extraction of information from within the RDF created by ModelDR and the original source within a single query. An example of this is the ingestion and subsequent addition of the data point RDF on a financial report in PDF format. In this example the PDF document is ingested in its native form and semantic RDF are in effect attached. These data points can subsequently be quickly resolved back to their source data – particularly useful with discovery-based analytics.

SUMMARY

Today's banks are finding it more difficult to establish effective data governance and quality assurance because the frequency of change to financial market data models has made it difficult for financial institutions to achieve harmonious, standardized views of data at a given point in time (often referred to as congruence).

Data point modeling is an effective means to solve the data architecture and aggregation challenges that hinder risk management, regulatory reporting, and effective decision-making. The problem is that data point modeling in a relational database still requires extensive amounts of ETL. By using a combination of RDF triples and document models, you can reduce overall data modeling by several months – if not years.

The MarkLogic database platform is unique in its ability to store both RDF and the original trade data from which the RDF is generated. The ModelDR solution enables the ingestion of information in its native form and subsequently exposes further information with triples. When combined, these complementary technologies ensure a highly agile solution that efficiently delivers regulatory reporting across a number of individual data models and sources.

To learn more about Enterprise NoSQL solutions for the Financial Services industry, please visit our website.

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