Aite PARTNER. ADVISOR. CATALYST.

Investment Bank Case Study: Leveraging MarkLogic for Records Retention and Investigation

MarkLogic

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

As regulatory changes and implementation continue to increase, one of the key challenges for global broker-dealers is their ability to capture and retain data from multiple sources to ensure the enablement of regulatory compliance and investigation processes. As these firms have operated typically in a siloed manner, driving growth through various regional and global acquisitions, changes in operational culture and processes are imperative.

This case study highlights a few of the challenges that a top-tier investment bank faced in terms of tying its data retention process to regulatory compliance and how it implemented MarkLogic's solution to address those issues. The archive project was multi-phased and the first phase was live after three months of implementation.

METHODOLOGY

This case study is based on in-person interviews with relevant executives at the investment bank and MarkLogic to examine how the bank implemented MarkLogic to address the bank's specific data retention issues.

CASE STUDY: TOP-TIER INVESTMENT BANK

The top-tier global investment bank identified in 2011 that it required a new archiving solution for the long-term retention of data from a wide range of operational systems across the firm. The focus was on application archiving and establishing a long-term data retention facility—a permanent archive of relevant business records necessary for investigations and regulatory responses. This specific archive currently only includes application data, as different facilities handle electronic communications and other data items.

The intent was to extract relevant business records from systems such as those for trading, client onboarding, and settlement processing and to allow for these records to be deleted from the operational systems in question. This would therefore restrict operational access to these records and free up storage for the 60 different operational systems in the scope of the project.

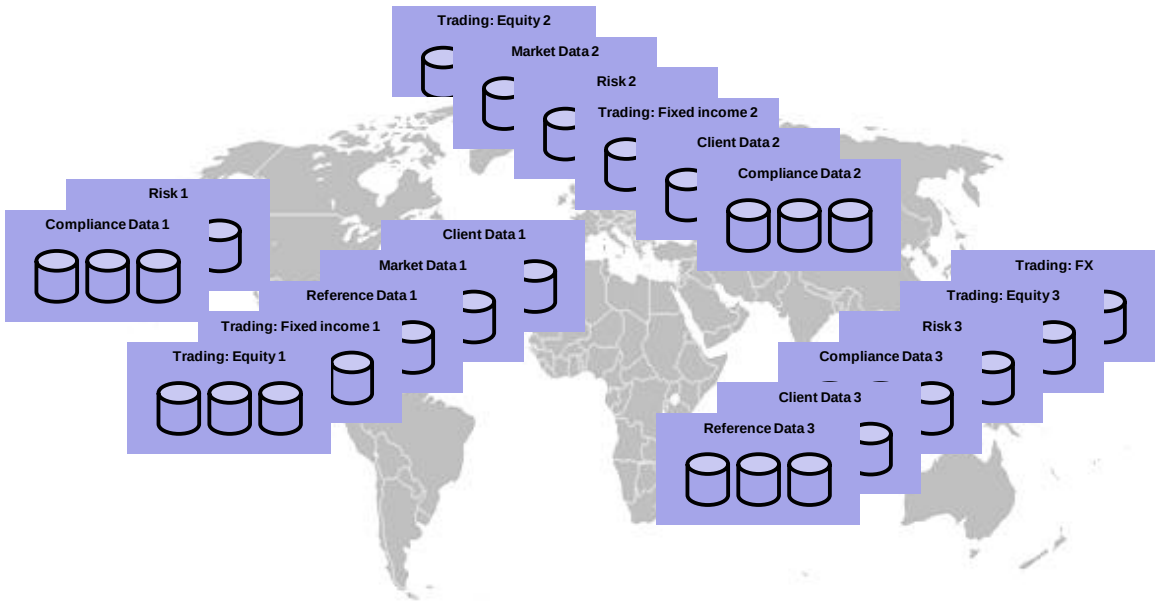
The project was kicked off with vendor selection in 2012, and the team revisited the timeline in 2013 when the scope and scale of the effort had become clearer. It is due to be completed in mid-2014.

PROJECT REQUIREMENTS

SOURCE SYSTEM VARIETY

The systems from which data had to be extracted were very different in terms of functionality, data formats, technology, and the business division in which they were situated. They ranged from very old Sybase servers and mainframes to very modern trading applications, and this high level of variety meant the systems landscape for the project was very complex. Accordingly, the data that had to be extracted was very wide and often densely packed as a result of the system variety.

Figure 1: A Complex Data Reality



Source: Aite Group

A FOCUS ON SPEED OF DELIVERY

The previous approach was to dedicate full-time employees from the books and records archiving team to transform data into a specific format and use a specific taxonomy before it could be onboarded onto a relational database system for archiving. The firm decided this approach would take too long due to the urgency of the project and the need to complete it within a 12-month time period. Speed of deployment was a key driver for the project, along with a desire not to rely heavily on in-house resources.

The project had to be carried out during a busy period when internal resources had been committed to other projects; hence there was a desire for minimized effort in dealing with source systems and the data extraction process. The archiving solution therefore had to be flexible in its ability to ingest data without requiring an extensive focus on extracting, transforming, and loading inbound data—it would need to be agnostic to the structure, shape, and size of this data.

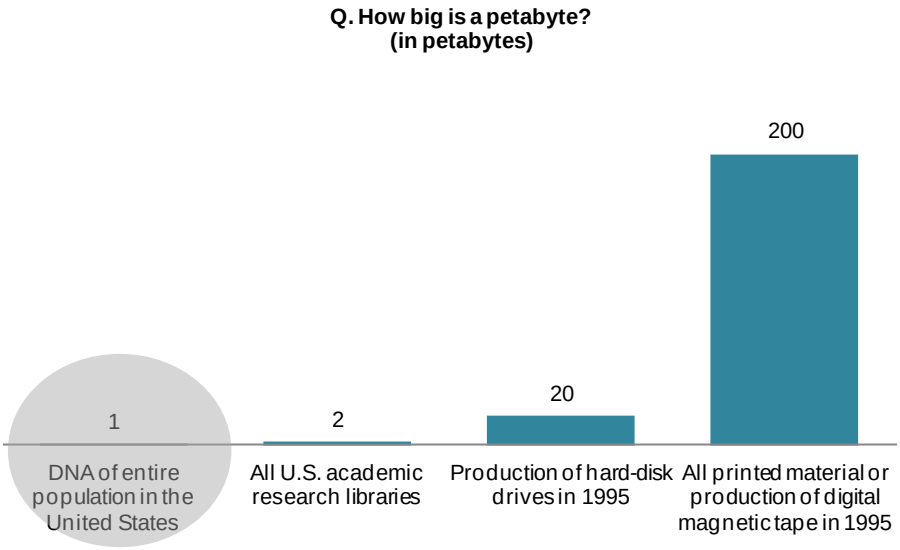
The firm opted for an XML-based solution and already had familiarity with the MarkLogic platform because it had been deployed in another area of the bank. This sped up the selection process because the information security legwork had already been completed for the other project.

SCALABILITY AND STORAGE REQUIREMENTS

During the scoping process, it was estimated that as much as a petabyte of data might need to be stored by the system, so the solution had to be large enough to scale to these requirements. One petabyte is 1,000,000,000,000,000 bytes of digital information and it is enough to store the DNA of the entire population of the United States (Figure 2). The solution also would have to

deal with multiple terabytes of data being ingested in one go during the data loading process. Once the data is in the system, however, it remains relatively static, with very little growth in the data being stored (unlike active transactional data, for example).

Figure 2: Relative Size of a Petabyte



Source: Caltech, Wikipedia

The piloting process taught the team that the decision to use an XML-based solution was sound and that MarkLogic's data-loading process would not restrict the amount of data that could be ingested at one time. It also allowed the team a high degree of freedom in its ability to support queries of different types—because of the way the data is stored and tagged, users are able to pose multiple types of queries.

QUERY SUPPORT

Currently, the firm's legal department poses a relatively low number of queries on a month-by-month basis, most of which are related to locating evidence for an internal or external compliance task. The team has a service-level agreement for delivery of the queried data within a five-day time frame. The firm has to do some level of work at the back end in order to assemble this data, but the high availability of the data simplifies this task overall. Data items can be linked together as required by the query due to the flexibility of the XML technology.

The team was not required to determine all of the future use cases for the data at the start of the process—unlike for a relational database project—because of the solution's thorough indexing process.

TIERED STORAGE

Though the firm is not currently taking advantage of the ability to tier its data, the team is aware of the potential to bring down data to lower tiers as it ages in order to benefit from cost savings.

The closer the data gets to its 10-year retention limit, the lower the requirement to keep the data in a higher tier because of the infrequency of required queries on this data. This need can be accommodated by MarkLogic's transparent (database-managed) tiered storage capability in version 7.

VENDOR IMPLEMENTATION AND SOLUTION

The implementation began with a proof of concept in 2012, and the onboarding of the application data began at the start of 2013. Thus far, around 40 applications have been integrated onto the platform, and the rest will be tackled before the end of Q2 2014.

IMPLEMENTATION CHALLENGES

The biggest challenges were faced in engineering the MarkLogic solution to work in the firm's environment due to specific jurisdictional regulatory requirements around permissioning, setup, and install. In order to ease this process, the firm had dedicated MarkLogic staff on site once a week.

In terms of technology, the firm faced challenges in building up the hardware horsepower required to support the data ingestion process. The MarkLogic solution responded well to being supported by a couple of large, dedicated memory-configured servers. On the internal system side, there was a challenge in getting some of the older systems to process data as usual at the same time they served up the data for ingestion.

The width of the data and the variety of formats was another conceptual challenge for the team, as it was initially difficult for the team members to understand how the XML technology would cope with this data. It was, for example, difficult to estimate how much memory would be required to support the technology due to the data's expansion during the ingestion process. Accordingly, the team overestimated the requirements at a petabyte of data, of which only 25% has been used thus far.

FUTURE ROADMAP

The firm will have finished the data ingestion process by Q2 2014, following the final stages of validation and data replication. The plan, however, is to kick off another similar project in 2015 and to factor in some of the lessons learned during this first project. One of the main lessons learned was the need to be more prescriptive about how the data is provided to the MarkLogic platform. Rather than attempting to extract data from aged systems and facing significant performance issues, the team will instead require end users to provide a copy of the data in question via a file-based method of delivery.

The 2015 project will involve around 100 systems and a similar volume of data from these systems as the previous project included.

MARKLOGIC

From its inception, the MarkLogic database platform for unstructured and semistructured data offered all of the features that come standard with an enterprise transactional relational database management system. Originally designed for document management for the publishing industry and government organizations, MarkLogic supports storage and retrieval of text, video, audio, images, XML, JavaScript Object Notation (JSON), and application-specific formats. Its newest release, MarkLogic 7, is a database platform with features such as enterprise search, Hadoop integration, and cloud-based deployment. MarkLogic 7 also has native semantic support, meaning it stores Resource Description Framework (RDF) triples and queries them using SPARQL.

BASIC FIRM AND PRODUCT INFORMATION

- **Headquarters**: San Carlos, California
- **Launched in**: 2001
- **Number of employees**: 280
- **Ownership**: Investors include Sequoia Capital, Tenaya Capital, and Northgate Capital
- **Main issues the vendor is trying to address**: Providing a big data/NoSQL database with enterprise-level features found in relational database management systems such as IBM DB2 and Oracle—including ACID (atomicity, consistency, isolation, durability) transactions, horizontal scaling, real-time indexing, high availability, disaster recovery, and user-level security authorizations and entitlements
- **Market positioning**: First and only enterprise-ready transactional, government-grade security NoSQL database
- **Key products and services**: MarkLogic Enterprise NoSQL database
- **Key statistics**:
  - More than 12 years in business
  - MarkLogic 7 released in 2013
- **Target customer base**: Users of unstructured and semistructured data
- **Number of clients**: More than 400
- **Current revenue sources**: Product licensing and consulting services
- **Implementation options**: Installed or cloud
- **Pricing structure**:
  - The Developer Edition license is free

- A license for the Essential Enterprise Edition with core functionality is US$0.99 per hour on Amazon Web Services, US$18,000 per year, or US$32,000 for a perpetual license
- Pricing for the full-featured Global Enterprise Edition is available on request; this edition supports distributed transactions, semantic indexing and search, geospatial data analysis, and tiered storage

KEY FEATURES AND FUNCTIONALITY

MarkLogic 7 adds the following features and functionality:

- Tiered storage
- Hadoop integration
- Cloud deployment
- Cluster monitoring
- Native semantic support, including RDF and SPARQL
- Built-in search
- Third-party authentication with LDAP/Kerberos

FOCUS OVER THE NEXT 12 TO 18 MONTHS

In addition to feature enhancements in the areas of semantics support and performance improvements, MarkLogic continues to expand beyond its core customer base of government and media organizations into financial services, insurance, telecommunications, and pharmaceuticals.

In financial services, MarkLogic has been used by major firms for the following use cases:

- Trade operational data store
- Reference data management
- Regulatory and legal compliance
- Customer analysis and insights
- Cybersecurity and fraud prevention
- Pre-trade decision support
- Information distribution

ABOUT AITE GROUP

Aite Group is an independent research and advisory firm focused on business, technology, and regulatory issues and their impact on the financial services industry. With expertise in banking, payments, securities & investments, and insurance, Aite Group's analysts deliver comprehensive, actionable advice to key market participants in financial services. Headquartered in Boston with a presence in Chicago, New York, San Francisco, London, and Milan, Aite Group works with its clients as a partner, advisor, and catalyst, challenging their basic assumptions and ensuring they remain at the forefront of industry trends.

AUTHOR INFORMATION

Virginie O'Shea
+44.(0)207.092.8129
voshea@aitegroup.com

CONTACT

For more information on research and consulting services, please contact:

Aite Group Sales
+1.617.338.6050
sales@aitegroup.com

For all press and conference inquiries, please contact:

Aite Group PR
+44.(0)207.092.8137
pr@aitegroup.com

For all other inquiries, please contact:

info@aitegroup.com