MarkLogic®

Universal Repository for Financial Services

MARKLOGIC WHITE PAPER · JANUARY 2022

Organizations have been creating value from simpler data for many years. But creating value from complex data can be much more difficult. Unlike simpler data, complex data requires metadata to create value.

Many organizations use MarkLogic to create a universal repository that connects complex data with any user seeking insights. By creating an agile response across disparate needs, organizations can vastly improve their speed and effectiveness in unlocking the value of complex data.

Current State Assessment

When dealing with complex data, many organizations try using tools that were designed for simpler data types. This often leads to significant effort and substandard results.

Below is a quick checklist.

	Yes	No
Are end-users frustrated when trying to work with complex data?	✓	✗
Is this difficulty causing specific business or operational issues?	✓	✗
Are there many different data types and sources that are of interest?	✓	✗
Do the data types include documents, emails, and other forms of text?	✓	✗
Are developers falling behind in keeping up with the rate of change: new requirements, new data types, etc.?	✓	✗
Is the lack of a shared environment that meets business requirements causing data marts and data lakes to proliferate?	✓	✗
Does maintaining security and compliance create challenges when sharing data between workgroups?	✓	✗

Contents

Introduction

Real-time information is the lifeblood of any financial institution. But what if you need to connect many different forms of information to serve customers and manage risk?

If a major customer leaves, is it clear why? A single 360-degree view of your customer across all product lines helps identify and remediate customer satisfaction issues.

Legal discovery is time-consuming and expensive. Generating required compliance reports is also time-consuming and expensive. Newer challenges – such as the "right to be forgotten" can create serious challenges when considering dozens – or maybe hundreds – of legacy systems.

Even core business processes – approving loans, detecting fraud – become easier and simpler when all relevant data is connected, wherever it was generated.

This white paper explores the difficult data challenges faced by firms in financial services and outlines a better approach. A "*universal repository*" ingests data in any form, connects it to other forms of data, and presents it in any way desired by users.

Organizations that have implemented a universal repository have discovered significant benefits:

- Faster access to relevant data insights for more users
- Accelerated application delivery by more developers
- Reduced risks through shared management
- Reduced effort in data sourcing, data transformation, data presentation, and system operations
- Reduced infrastructure costs

The Challenges

Data Is in Silos

Information is created one application at a time, which results in data silos. Multiple data silos force additional effort from anyone who needs information across multiple applications. Even answering a seemingly simple question can involve days (or weeks) of querying, transforming, and assembling the data into the needed form.

As more applications come into the environment – and more users need information that spans applications – the inefficiencies rise exponentially.

Production systems get hammered with users querying – and re-querying – performance-sensitive applications. Shared network bandwidth is consumed with data transfer after data transfer. Data marts proliferate, with attendant costs and risks. A complex world of applications has now been augmented with a complex world of data marts and other ad-hoc repositories.

Each has its own perspective on "truth" which is not shared across others, which creates additional challenges.

This decentralized approach also creates serious challenges when implementing new organizational-wide mandates.

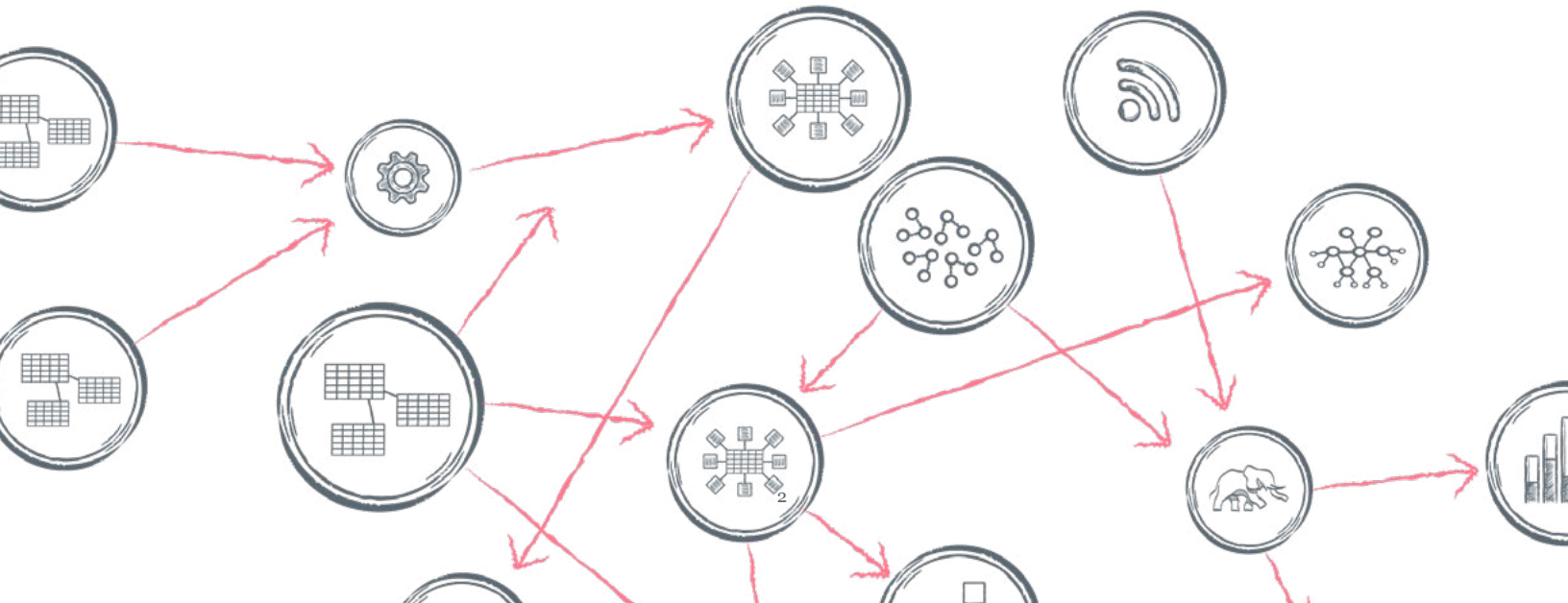
Data Is Complex

Customer data can be inherently complex. There are transactions, payments, support issues, survey results, and more. Beyond this, there is information providing context: organization of staff within a customer, relationships between customers, relationships within a customer, … and much more.

Firms may maintain relationships with individual customers for decades. Individual transactions including bonds, mortgages, and some derivatives may be in operation just as long. The laws governing these transactions vary from year to year and state to state. As requirements change, the structure of the data representing them changes as well.

And, at any time, a firm may need to pull together all the documents relating to a set of transactions to respond to a subpoena or other reason.

Many of the IT infrastructures used to originate, process, and securitize financial transactions were developed when the industry was simpler. During that period, there was little need to have a consolidated view of all aspects of a transaction's complete lifespan for both risk mitigation and regulatory compliance. In fact, it often made sense to build a variety of siloed systems – as they could be developed faster and more easily by creating them independent of each other.



New Data Is Hard to Integrate

Because legacy systems were often built on relational databases, which require fixed and rigid schemas, it can be difficult to integrate large numbers of existing – and even new – data sources. The ETL effort needed to bring the data into conformance with existing relational schema may require months of effort and several days to load each new data set.

With different systems containing different types of documents, it can be extremely difficult to query all the documents as an integrated whole. Querying across multiple data stores chews up processing which can disrupt production systems.

Efforts to duplicate data are often done with little attention to governance. This leads to inconsistent data sets, with each replica having its own update policies and each replica doing its own transformation and enrichment of the data.

This jungle of data is difficult to secure (sometimes it may even exist on spreadsheets) even though it contains critical customer information.

Provenance Must Be Tracked and Easily Available

Understanding how data values were derived is a critical business function. If data provenance cannot be determined, it cannot be verified that the data is accurate. Firms that cannot demonstrate provenance risk bad decisions, fines, and other sanctions.

The Business Is at Risk

These technical, data-related issues lead to a variety of business challenges. These include:

- Data quality issues
- Increased risks
- Limited analytics
- Slow and inferior decisions
- High costs
- Limited ability to detect fraud
- Inferior pricing of securitized instruments

Implementation of a Universal Repository can help organizations in the overcome these challenges and risks, by helping you handle the potentially thousands of different variants of documents that causes headaches in topics like analytics, transaction processing, origination, and securitization.

Overcoming the Challenges

How do you move from the manual processes, errors, and inefficiencies of today's overlapping and unmanageable data silos to a single, unified golden source repository? How do you do it while keeping your business running and protecting possibly decades of investment and work?

To get past these limitations, it is necessary to build an infrastructure that includes:

- A universal repository that is a single source of truth with a consolidated view of ALL information
- Quick access to new data and new data sources – minimal ETL
- Support for new data governance and regulatory demands
- Enterprise data integrity, reliability, and security
- A complete history of each transaction
- Ability to handle many billions of documents
- Geospatial queries

Combining all these into a universal repository is not a utopian fantasy. This white paper is based on work being done at major banking institutions, and shows how a new-generation system should work.

Building the Solution

Managing Metadata & Documents

The biggest issue in building a repository is the number of different types of data involved. For example, in tracking a mortgage over its possibly 30-year lifespan, there are documents on customer information, checking and savings statements, check images, credit reports, and mortgage applications. The formats of many of these documents vary from state to state or possibly even from company to company. The versions of the documents used today may look very different than the versions that existed in 1995 or 2005 – and it is necessary to store them all. Every time regulations change, new versions of many of the documents will be created.

A large bank that originates and processes millions of transactions will store many variants of different document types – each of which may have a different implied schema. In addition to mortgages, customers will often conduct a wide variety of other transactions with the bank as will.

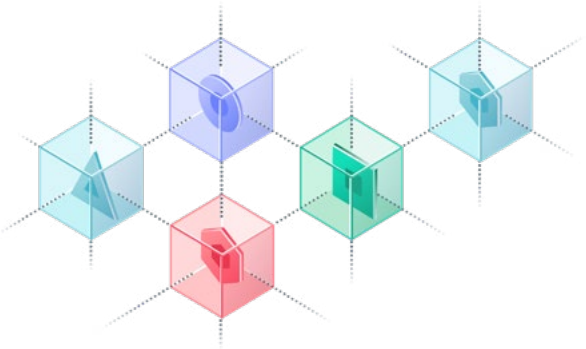
In some cases, a firm may not want a complete document repository – where the actual documents are centrally stored and queried. Organizations often prefer to keep many of the actual documents with the primary systems that create and maintain them. Instead of creating a document repository, a metadata repository is maintained. This makes it possible to determine which documents meet a set of search or query parameters and quickly locate and retrieve them from the primary owner's systems.

As time has passed, firms have gradually increased the metadata stored on individual documents – which often makes it possible to answer some questions with just the metadata.

Metadata repositories can be easier to build than document repositories as their schemas can be simpler.

However, even with a pure metadata repository, there is still a lot of work to do. To start with, there are many different business processes involved in the life of a transaction, including check processing, monthly statement creation, and origination processing. In fact, an infrastructure may contain dozens of primary systems. These systems have largely been developed independently of each other, and the metadata varies from system to system. Mergers and acquisitions mean that different metadata approaches coexist in a single firm. These days, after origination, long-lived transactions are regularly bought and sold – so documents, and perhaps metadata, created by many different firms may coexist in the same database. So, even with a metadata repository there can still be thousands of variants of document types coexisting within a firm. In many cases it will be desirable to create mixed systems with metadata mixed with the full data for important and commonly accessed data subsets.

Trying to pull all these primary or metadata documents together with ETL and a relational database is a recipe for disaster. The large existing inventory of document types, and rates at which new document types and variants of existing types grow, means that any attempt to engage in a systematic modeling and ETL effort will likely never finish.



Loading, Modeling, and Harmonizing the Data

Using the MarkLogic® Data Hub Platform as the technology for the universal repository provides all the functionality needed to load, optimize, query, secure, and manage your mortgage metadata and documents.

Data Load

The first requirement in building a document or metadata repository is to just load the data. Instead of defining a schema and then engaging in extensive ETL to force underlying data sources to fit into that schema, with MarkLogic you just load the data "as is."

As it is being ingested, MarkLogic's Universal "ask anything" Index makes it immediately available for searches without any ETL required. Additionally, structured queries can be performed against the existing metadata in the primary PDF, Word, and other documents, or against the XML descriptors or JSON tags found in metadata documents – again without any modeling or transformations.

This ability to powerfully access the data on load means that the repository can provide value from day one. Users can search and query and get results faster and more accurately, even before more advanced development begins.

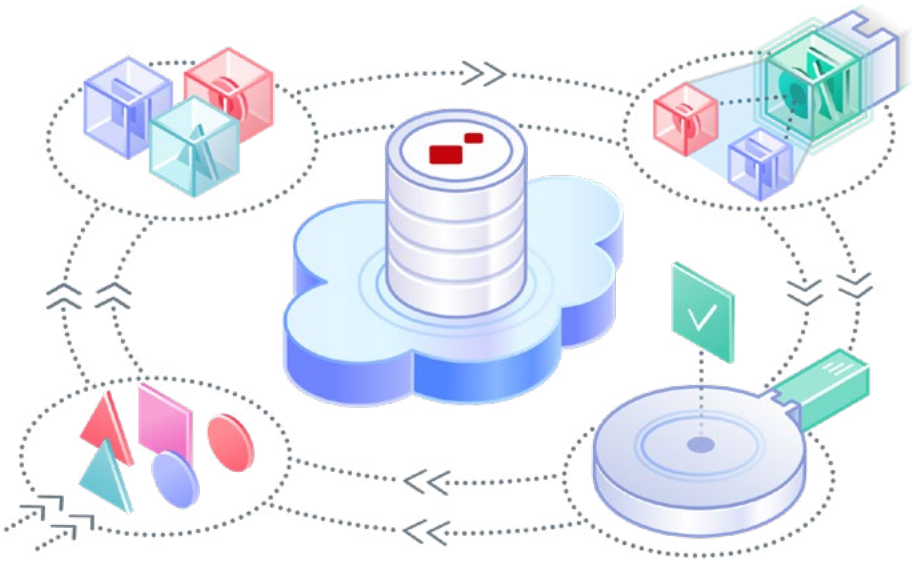
Modeling

The number of types and complexity of the data sources generally make fully modeling and harmonizing the data an impossible task – data modeling and ETL cause many big data projects to fail.

MarkLogic makes modeling manageable with agile data modeling. With agile data modeling and harmonization, you model and harmonize just what you need for your immediate deliverables. As time passes, you enhance the model – with the changes you make to it reflected in both new and existing data. When you need to enhance or change your data model, the changes are applied to your existing data as well as new data. This is possible because everything you need is still in the document.

This approach is very different from most products in the market and is key to why MarkLogic customers are successful, especially in areas with changing requirements, like regulatory compliance.

MarkLogic's technology for implementing agile data modeling and harmonization is the Data Hub Platform. The Data Hub Platform provides



a graphical environment for loading diverse data, creating and applying data models which span relational and semi-structured XML and JSON data, and which allows all types of data to be queried and searched as an integrated whole.

MarkLogic provides a complete framework for integrating and making complex and diverse data sets accessible.

This includes data duplication, alerting, automatic tracking on lineage, advanced semantics to make self-service access to complex data sets possible to business users, and much more.

Provenance

MarkLogic offers superb provenance and lineage capabilities. This is in part because MarkLogic stores both the original "as-is" data and the derived curated data in the same underlying document. In addition, as part of the data modeling process, MarkLogic makes a record of transformations done to the data. This record is based on the PROV-O standard.

With this, especially if transformations are done inside of the Data Hub, then it will always be possible to do forensics on any data value and determine how it came to be.

What did you know and when did you know it? – An important issue in today's mortgage world is for firms to have the ability to show regulators and others what they knew and when they knew it. Bitemporal functionality is the ability to show what the state of a database was at a point in the past.

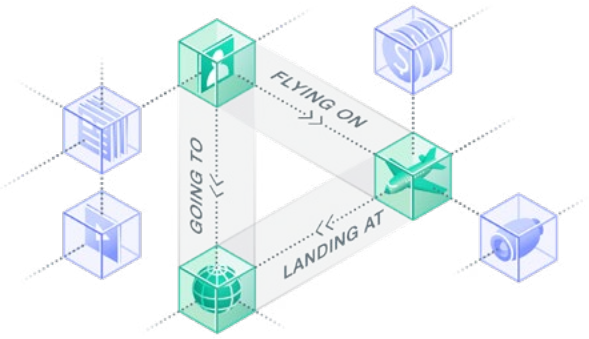
With relational technology, it is especially difficult to implement bitemporality with long-lived instruments like mortgages or other complex transactions. Document structures used in mortgage processing and securitization are continually changing which, in a relational-based system, requires schema changes. Displaying data with their schemas as it existed 10 or 20 years ago is a nearly impossible job for a relational database.

Accessing the Repository

As we have discussed, structured queries can be performed against MarkLogic without any data modeling or ETL.

MarkLogic includes a world class search engine; word/phrase search, rankings, facets, auto-completion, wild card searching, stemming are all easily available.

When coupled with the structured queries, the MarkLogic Universal Index provides powerful tools for data discovery and self-service access to data, even data that has not yet been modeled. On top of these is the integration with semantics and SQL.



Semantically Enhanced Queries

When a user is faced with TBs or PBs of diverse and inconsistently formatted data, it can be an almost impossible task to even understand what is available, much less be able to isolate and extract the needed data.

An important element to bringing order to these data sets is using semantics to enable data discovery on top of document and SQL data. By adding a firm's knowledge of itself and its industry to MarkLogic in the form of a semantic ontology, MarkLogic can become a smart data platform that can expand searches and queries with domain knowledge.

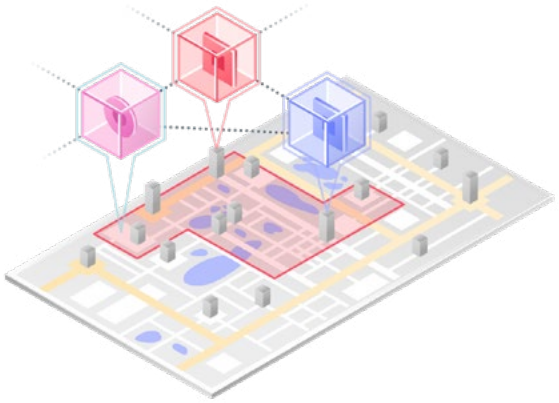
For example, if a user queries on 'bonds' and 'high yield' they will often not get results they expect because much of the data in the system is indexed with different terms than the user specified. A Muni is a kind of a bond. High risk is another way of saying high yield. In many systems this type of mismatch causes users to miss a large majority of the data they are looking for. MarkLogic's multi-model approach makes it easy to embed and use the firm's domain knowledge and let users get the data they need without the assistance of IT.

In addition to semantic ontologies, semantic triples can link related content. In relational databases, foreign keys link records, making it easy to perform queries like 'show me all the orders for a customer.' Many document-oriented technologies struggle with this kind of query. Using semantic triples, this kind of query can easily be done with any type of data.

With these two semantic approaches data becomes far more manageable.

Location-Based Queries

Geospatial indexes make it easy to do location-based queries. For example, if you are pricing a basket of mortgages and want your pricing algorithm to take into account that properties near a specific location may have been affected by a recent oil spill, geospatial searches make it easy.



Enterprise Readiness

Mortgage document and metadata repositories are very demanding in terms of data complexity, scalability, security, and data reliability. They often contain billions of documents with widely varying, implied schemas. While there are many technologies that can handle one or two of these demands, there are few that can handle all of them.

- Relational-based approaches are known for their security and data reliability, but they have great difficulty handling the required levels of complexity and scale.
- Newer NoSQL and Hadoop technologies can often handle the scale, and are also often better than relational approaches in handling complexity – although querying the wide variety of data types (document, geospatial, etc.) found in a modern repository can require combining multiple technologies, and extensive development and design before queries can be performed. And, they generally lack the type of security and data reliability organizations require and have come to expect.

At MarkLogic, we believe that an enterprise-ready technology must effectively meet all of these requirements of handling data complexity, scalability, security, and data reliability. Below you will learn how MarkLogic is differentiated from other NoSQL and Hadoop technologies in terms of security and data integrity.

Security

Security is vital in any system that contains customer PII (personally identifiable information) data. PII data is both protected by privacy laws and, when hacked or exposed, can result in the firm being on the front page of major newspapers and other media in very uncomplimentary ways. In recent years there have been many major breaches of customer data, which have been extremely damaging to the firms involved.

To have a fully secure system requires encryption, user authentication, and document and cell-level security. Old-school relational-based approaches can generally handle these kinds of security requirements. However, security and reliability are where many of the new NoSQL and Hadoop-based technologies fall down.

The database with the best security of all the NoSQL databases is MarkLogic. MarkLogic is the only NoSQL database with NIAP Common Criteria certification. It is extensively used by intelligence agencies. In fact, MarkLogic has had government-grade security from the start.[1]

Data Integrity

If you are storing billions of documents in a repository, you need to be sure that all the data that is supposed to be there is there. Providing inaccurate responses to subpoenas or customer inquiries can lead to fines and damaged reputations. The basic key to data integrity is ACID (Atomicity, Consistency, Isolation, Durability) transactions. ACID is the key to transactional consistency and data integrity. With ACID, if a database claims a record is committed, it is.

MarkLogic has had ACID transactions since it was first released. In fact, it is not even possible to turn ACID off. This is not the case with most other products in the NoSQL and Hadoop space.

If you want to be sure your queries return accurate results and you get the same answer when accessing different machines, then you need ACID transactions. Fortunately for our clients, MarkLogic's data consistency guarantees, high availability, and disaster recovery all operate at the enterprise level of performance needed by a modern mortgage document or metadata repository.

Lower Costs

A universal repository can substantially mitigate the cost of document review by providing a centralized store of metadata on the underlying documents and other data. Attorneys can quickly select or reject documents by scanning metadata on large numbers of documents at once instead of individually examining each one.

Meet Today's Data Challenges

Building a universal document or metadata repository has been nearly impossible with relational technologies, but major banking institutions are now succeeding with MarkLogic – with enterprise-level security, high availability and disaster recovery, ACID transactions, in a clustered environment that can scale to many billions of documents.

The following pages show examples of how a universal repository based on MarkLogic works in practice, for a variety of financial services use cases.

1 For a more in-depth look at MarkLogic security see https://www.marklogic.com/product/marklogic-database-overview/security/

Universal Repository in Practice

Legal Discovery

Finding documents to respond to a subpoena is like finding needles in a haystack.

As any firm involved in mortgage processing, servicing, or securitization can tell you, legal discovery and subpoena response is a:

- Major cost of business
- Potential risk to reputation
- Potential source of legal sanctions

Properly responding is hard since it can encompass all the data produced by a firm. This can include, among other things, metadata, documents, databases, spreadsheets, "deleted" data, legacy data, emails, text messages, and instant messages – all of which are stored in various storage systems, including archival systems, including tape, microfiche, and more.

A very small subset of examples of the kind of information that can be demanded in discovery includes:

- Copies of every document signed by a specific officer of the company
- Copies of any documentation that allows for a given employee to sign documents
- All manuals on the drafting or execution of foreclosure and mortgage-related documents
- A list of all employees, dates of hire and termination, and their duties
- All documents regarding any contracts with a specific law firm
- Any price lists provided to prospective customers
- All communications with a specific firm relating property-related transactional documents or records filed in any court
- All financial transactions with a specific firm

Pulling this diverse data together from dozens or maybe hundreds of separate data sources can be a nightmare.

Legal sanctions[2] provide a strong incentive for companies to create comprehensive document retention and retrieval systems before litigation is anticipated, as failure to respond in a timely manner can result in summary judgments or instructions from the judge to assume missing responses are unfavorable to the firm.

In order to effectively respond to discovery, a firm's response team must be thoroughly familiar with:

- What type of relevant information exists
- The format in which the information exists
- Where the information is stored
- How accessible the information is

A universal document or metadata repository greatly simplifies this task.

2 For example, Rule 37 – Failure to Make Disclosures or to Cooperate in Discovery; Sanctions of the Federal Rules of Civil Procedure

Generating Insights

Generating insights on your customer's concerns, plans, pain points, and likely behavior is critical to keeping existing and attracting new customers, and expanding your relationship with them.

The data available for generating insights is more diverse and varied than ever before.

Insights can be gained from social media, trading data, news alerts, and many other sources. Web crawling can make collecting valuable information much easier.

From an IT perspective this means you need to have a platform that can handle virtually any kind of data and pull the different types together into a single unified whole.

80% of data scientists and other research creators' time is spent on obtaining and preparing data. A key reason for this is that in the early stages of analysis, data scientists do hypothesis testing where they try out new ideas – and if they fail, move on to the next one.

The one thing you do not want in this type of situation is to have to do time-consuming formal data modeling and ETL just to see if an idea works.

You want to be able to get high-quality data with minimal effort, test the ideas, and then, if you have good results then do the work needed to productionize it.

With a MarkLogic universal repository, semantic ontology query expansion integrated with SQL, search, document filtering, and geospatial make data easy to find. Agile data modeling and harmonization reduce time spent on processing the data to just what is needed for immediate needs.

Customer Service, Reporting, Alerting, & Regulatory Compliance

High-quality customer service is critical. Having an infrastructure that allows firms to quickly evaluate, respond to, and resolve customer complaints is essential to keeping costs low, maintaining a good reputation, and growing.

A universal repository allows your firm to handle issues like analyzing and responding to customer assertions of error, providing accurate legal filings for foreclosures, and understanding when events (natural disasters, major layoffs, etc.) occur that impact your consumers' ability to meet their obligations.

Regulatory reporting has many of the same technology needs as customer service.

It takes many forms. Firms have reporting responsibility in areas like their exposure to default risk of other firms, Suspicious Activity, Anti-Money Laundering, and many other industry-specific requirements.

Customer service and compliance often cut across business lines and require a wide variety of data, both structured and semi-structured.

The diversity of these needs demands equally powerful technology. All kinds of data need to be available. The data needs to be easily accessed without endless processing. Much of the data is sensitive and requires high levels of security. It is often necessary to understand the history of the data.

A central repository of all the firm's relevant data sources makes it far easier to pull together the information needed to generate compliance reports. A prime need is the ability to quickly and accurately generate reports and issue alerts when events occur that impact existing customers.

For compliance especially, data provenance is not negotiable. Not being able to produce it can result in fines or other sanctions.

Providing lineage and provenance is difficult for many technologies. In many systems data may go through multiple ETL steps before being presented to users. Once the data arrives at its final destination, it can be difficult to determine the source of individual data elements.

MarkLogic largely solves this problem. The MarkLogic Data Hub Platform stores, in the same document, the original as-is data (it is never deleted), the final curated (golden) data, and the platform can be configured to store a record of all transformations performed on the data.

With all this it should always be possible to do the forensics needed to determine where individual data elements came from.

To enable easy access to complex customer and regulatory data it is necessary to combine a variety of search and query approaches including SQL, search, document filtering, and semantics. SQL provides the structured access and aggregates customer service use cases require. Search and document query approaches provide flexible filter capabilities which are powerful and more flexible than SQL. Semantic ontologies and semantic links between records and documents allow the repository to understand its data and the relationships within it, so that users can access it without having to become experts in how the data is laid out and indexed.

MarkLogic provides the unique benefit of providing all these data access technologies in a single platform. MarkLogic stores all the underlying data types in a single data store and allows queries to simultaneously incorporate as many access approaches as needed by the use case.

Trade Repository

In the past, trading desks maintained their own data and handled their own reporting. Today, risk management, reporting, and sometimes trade processing occur on a firmwide basis.

The diversity, complexity, and changing nature of trading and the regulatory demands on it make it difficult to provide a unified understanding of a firm's positions, risks, and opportunities.

The path from quote request to regulatory reporting is long and complicated.

Each instrument type has its own set of systems, and each trade must pass through multiple systems to handle quote responses, trade validation & capture, reconciliation, and more. Each system has its own data store, with its own data structures.

All of this makes trade data complicated.

With a traditional relational-based approach every column of every row of every table may need to be modeled and ETL developed for it before development can start. This can literally add years to development.

What are some of the core capabilities of a first-class trade hub?

The ability to grow in an iterative, agile, way. Powerful search and query capabilities. For real-time systems you may need ACID transactions.

The key to an iterative approach is agile data modeling and harmonization. To fully model every aspect of each derivative and complex message type in the repository is a prohibitive task that would add years to the building of the repository, and which would be out of date before it was built. MarkLogic's Data Hub Platform provides a comprehensive approach to building an agile data repository.

To quickly onboard new derivative or SWIFT message types, agile modeling and harmonization are critical.

To enable easy access to complex trade data it is necessary to combine a variety of search and query approaches including SQL, search, document filtering, and semantics. SQL provides the structured access and aggregates trade repository use cases require. Search and document query approaches provide flexible filter capabilities which are powerful and more flexible than SQL. Semantic ontologies and semantics links between records and documents allow the repository to understand its data and the relationships within it so that users can access it without having to become experts in how the data is laid out and indexed.

MarkLogic provides the unique benefit of providing all these data access technologies in a single platform. MarkLogic stores all the underlying data types in a single data store and allows queries to simultaneously incorporate and many access approaches as needed by the use case.

When the trade repository will be accessed for operational needs, ACID transactions become vital. Without ACID, queries may give incorrect results because only parts of multi-record transactions have been committed when the query is performed. Without ACID, the risk of data corruption grows for the same reason.

About MarkLogic

MarkLogic helps customers create value from complex data faster. Our platform ingests data from any source, creating and refining metadata to support powerful models. Customers use these models for deep search and query, building enterprise applications, and bringing unique insights to analytics and machine learning.

For more information, please visit www.marklogic.com.

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