MarkLogic®

SOLUTIONS

The New Generation Operational Data Warehouse

Making Analysis Operational

For years the Enterprise Data Warehouse (EDW) has been the driving force behind data analytics. Based on relational database management system (RDBMS) technology and dimensional models such as the star schema, the EDW was conceived to provide a long-timeline cross-organizational view of the enterprise for analytical purposes.

As the need to analyze more real-time cross-organizational data emerged, a complementary repository – the Operational Data Store (ODS) – was conceived to work in tandem with an EDW. Over the years as technology has evolved, the line between the EDW and ODS has blurred into a concept of an Operational Data Warehouse (ODW), allowing for data warehouses to cover a more complete view of data from a time and timeliness perspective.

Despite these improvements, however, most Operational Data Warehouse implementations are RDBMS-centric. This means the vast majority of today's implementations are limited to analysis of only highly structured data – and the population of the warehouse is dependent on brittle extract-transform-load (ETL) processes, characteristics which limit agility and time to delivery.

In contrast, an ODW built on the MarkLogic® Enterprise NoSQL platform not only improves upon the traditional capabilities associated with an ODW (e.g. ingesting large amounts of data and making it available for query in real-time), but also makes this capability available for a wider variety of data, in a much more agile way than previously possible. MarkLogic provides several key advantages in an ODW implementation:

- Data is stored as-is without the need to conform to a single schema on write
- MarkLogic handles not only structured data, but also unstructured and graph data in the same data store, all at enterprise scale
- Once ingested, all data – regardless of shape – is indexed and immediately searchable by the powerful built-in search engine
- The BI analyst can take an **active** role in evolving the knowledge base

Background

To better understand how MarkLogic offers a superior ODW platform, it is first helpful to examine the limitations of traditional data warehousing designs, and how they fall short for today's Big Data needs.

The Cost of Prerequisite Modeling and ETL

Data warehousing typically relies on a data pipeline that consolidates information from multiple downstream OLTP systems to create a consolidated view for the purpose of data analysis. With RDBMS-based data warehousing, a consolidated data view also requires that an all-encompassing – or *canonical* – data model be created to house the data. Creating this canonical data model is highly dependent on both the **number** and **variety** of downstream systems. The more systems and the greater the variety of data, the more difficult it is to create a canonical model.

In addition to the prerequisite modeling dependency, there is also a dependency on resource-intensive processes to extract, transform, and load (ETL) data from the source systems into the data warehouse's canonical data model. The degree of difficulty of this process is also directly dependent on the number and variety of downstream systems:

- For each downstream system, one must write a new (or revisit an existing) ETL routine
- Each ETL routine, whether it is hand-coded or facilitated by a framework or tool, is subject to a software development lifecycle (SDLC) release process and associated QA and testing

ETL processes also introduce business challenges in assessing the quality and lineage (or provenance) of the data in the warehouse. Since the transformation logic is managed completely separately from the transformed data that is stored in the warehouse, analysts and IT staff need to have visibility into the logic in the ETL tool whenever questions arise – or documentation is required – about the veracity of any data values or derivations.

Figure 1: A conceptual diagram of an EDW process; gears in the diagram represent ETL, and the grids represent schema-revisits

Beyond the burden of these ETL processes and schema-revisits is the impact of the entire SDLC lifecycle on the discovery process. New insights from an organization's data are very dependent on this lifecycle, whether it is the initial release, or later adjustments to the data warehouse. This results in something akin to a waterfall development lifecycle where new discovery is highly dependent on prerequisite modeling and often time-consuming technology delivery.

Today's data needs have evolved to the point that these dependencies are no longer sustainable.

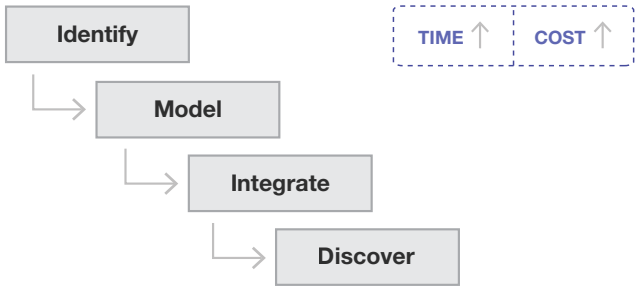

Figure 2: Waterfall development increases time and cost

MarkLogic®

MarkLogic and the Discovery/Model Loop

Using MarkLogic for an ODW solution minimizes or eliminates these dependencies. MarkLogic has a schema-flexible approach to storing data, which reduces or eliminates the need for complex ETL. With its comprehensive indexes and built-in search engine, MarkLogic is able to ingest data of any shape (e.g. structured, unstructured, graph) and provide robust insights and discovery capability immediately upon ingestion, without having to first conceive of the perfect data model. And while it is certainly true that models are a necessary and important part of any analysis endeavor, what's not necessary or practical anymore is the need to design a perfect model up-front before any analysis can be done.

Instead, MarkLogic supports a discovery/model loop, where a model evolves alongside the discovery process in a very fluid and agile way.

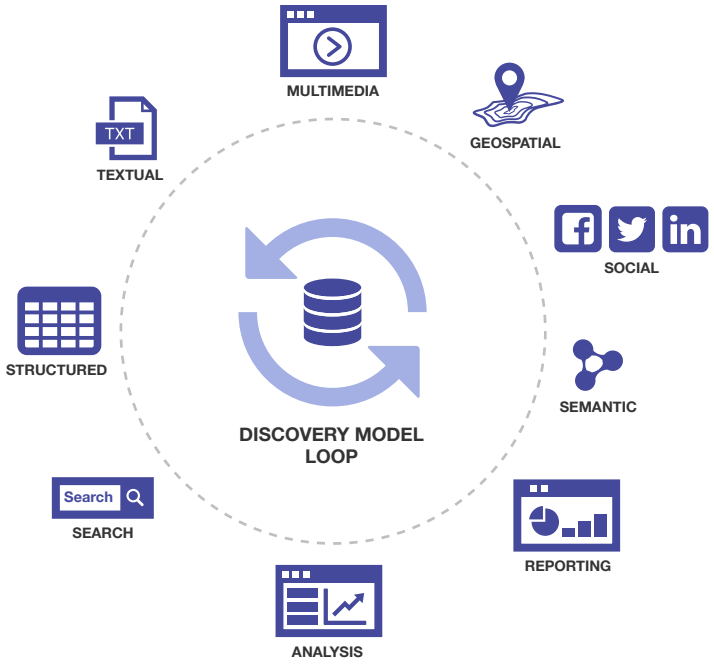

Figure 3: MarkLogic lets you evolve the model alongside the discovery process

Enterprises with a Hadoop investment may be somewhat familiar with this concept through the use of HDFS and map-reduce. The difference with MarkLogic is that the discovery/model agility comes from a real-time enterprise-grade database with ACID transactions, security, and other important operational characteristics built in. MarkLogic also allows organizations to leverage their existing Hadoop investment by providing Hadoop integration features that complement what's possible, resulting in agility for all types of analysis workloads.

From Read-only Analysis to a Discovery Conversation

For most, data analysis is largely considered a read-only activity – but imagine if you could have a contextual conversation with your Data Warehouse, instead of merely asking questions. With MarkLogic, the discovery process takes on another dimension: MarkLogic provides a foundation for read+write analysis where the people doing the analysis can play an active role in evolving the corpus of knowledge. A business intelligence (BI) analyst can directly enrich the data warehouse with human-derived insights, in context, in real time and without having to rely on a rigid data pipeline.

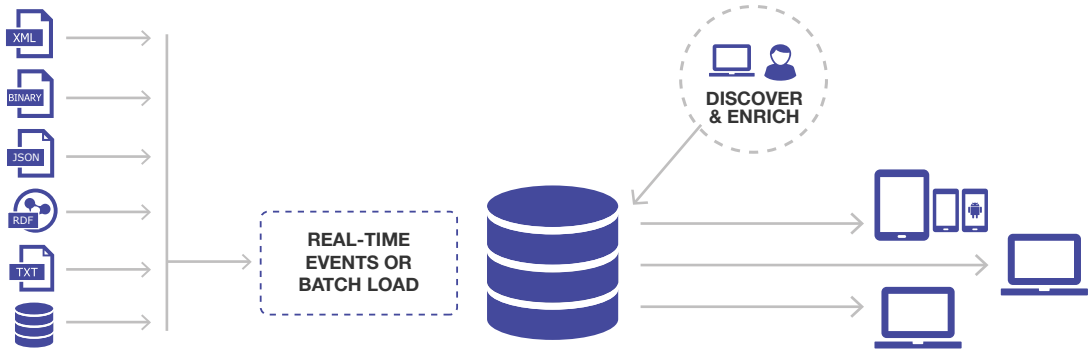

Figure 4: Conceptual diagram of an Operational Data Warehouse

One of the key enablers of this interactive two-way analysis capability is MarkLogic's robust support for Semantics. MarkLogic is – amongst other things – a first-class Semantics triple store, at a level of capability aligned with our Enterprise pedigree. With support for various W3C Semantics standards – RDF, SPARQL, including inferencing and update, to name a few – alongside our other Enterprise features such as ACID support and fine-grained security and auditability, BI analysts can assert new facts into the data warehouse for use by other analysts, all while making inferences from downstream warehouse data and from the asserted facts of others.

Imagine if you could have a contextual conversation with your Data Warehouse, instead of merely asking questions. With MarkLogic, a BI analyst can directly enrich the data warehouse with human-derived insights, in context, in real time and without having to rely on a rigid data pipeline.

Operationalizing Analytical Insights

Most data warehouse deployments are purely "downstream" systems – aggregating data for use in analytical activities. Turning the results of this analysis into operational activities is a separate step, performed by separate processes and systems – there is a built-in lag between analysis and operational impact. MarkLogic, however, lets your business react in real-time to analytical insights. MarkLogic was architected from the ground up to handle simultaneous read and write workloads at scale, and also includes built-in alerting capabilities. This means you can automatically fire off any number of tasks or notifications to the appropriate systems or people, in real-time, based on parameters you define. One MarkLogic customer uses over 450,000 different alerts to help power business operations.

Conclusion

Organizations around the world are using MarkLogic to create a new breed of Operational Data Warehouse for mission-critical environments. They rely on MarkLogic's agility to ingest and index data as-is, its powerful Semantics capabilities, and its tested enterprise features. And they benefit from the MarkLogic transactional database platform that not only lets them handle mixed analytical workloads at scale, but also delivers a foundation for two-way analytics, while drastically simplifying what it takes to make post-analysis activities operational.

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