White Paper
MarkLogic and Intel for Federal, State, and Local Agencies

Simplifying Data Governance and Accelerating Real-time Big Data Analysis for Government Institutions with MarkLogic Server and Intel

Reduce risk and speed time to value using an integrated NoSQL + Big Data solution built on MarkLogic and Intel

Executive Summary

The era of Big Data is driving considerable changes in how governmental agencies manage and use the varied types of data they acquire and store. The legacy Relational Database Management System (RDBMS), Enterprise Data Warehouse (EDW), and Storage Area Network (SAN) infrastructure used by agencies today to create siloed data environments is too rigid to accommodate the demands for massive storage and analyses on a larger and wider variety of data. Forcing this legacy architecture into today's public sector organizations requirements is costly and risky.

MarkLogic has integrated their new-generation Enterprise NoSQL platform with Apache Hadoop* optimized for Intel® Architecture to deliver a powerful platform that provides all the features of MarkLogic and Apache Hadoop with the data governance and security IT departments need. Running on Intel® technology and the enhancements Intel has brought to Apache Hadoop, this integration gives public agencies a true enterprise-class Big Data solution with government-grade security for storage, real-time queries, and analysis of all their data.

This paper summarizes the issues public agencies face today with legacy RDBMS + SAN data environments and why the combination of MarkLogic, Apache Hadoop, and Intel provides a government-grade solution for Big Data.

"MarkLogic has integrated their new-generation Enterprise NoSQL platform with Apache Hadoop* optimized for Intel® Architecture to deliver a powerful platform that provides all the features of MarkLogic and Apache Hadoop with the data governance and security IT departments need."

Table of Contents

MarkLogic, Hadoop, and Intel in Federal, State, and Local Agencies

All government customers want to reduce costs and often look to open source solutions to help with this effort. Many gravitate towards integrating Hadoop as part of their enterprise architectures in order to deal with their vast volumes of data, but Hadoop is not the entire solution. Coupled with MarkLogic, however, it becomes a solution to cost-effectively integrating and storing unstructured, semi-structured, and multi-structured data. The MarkLogic, Hadoop and Intel combination provides customers with tiered storage that offers secure, fast, and reliable access to data over its complete lifecycle. For these reasons, government agencies choose MarkLogic over other solutions.

Government agencies are the original "Big Data" creators and users, and they have stringent requirements for data management – including extensive security, auditing, and retention policies. They also need to be effective stewards of taxpayer dollars. By combining MarkLogic with Hadoop and Intel Architecture, government organizations can cost-effectively integrate, store, search, share, and archive their data as needed to ensure mission success.

For example:

- Using MarkLogic with Hadoop on Intel Architecture, federal healthcare organizations can more easily ensure patient eligibility and prevent fraud waste and abuse. By easily consolidating information across silos, such as data on providers, patients, procedures, watch lists, etc., MarkLogic can generate real-time risk alerts, while Hadoop enhances overall scale and data logistics capabilities. Together, the system provides operational efficiency to the overall process.
- Using MarkLogic with Hadoop on Intel Architecture, intelligence and law enforcement agencies can extend not only their ability to respond to greater volume, but also better incorporate their unstructured data into the analytical process. Previously, such data has either been too time-consuming to incorporate or simply not possible with traditional technologies.
- In the all-source analytical space, MarkLogic's Semantics triple store, paired with the Apache Hadoop and Intel, provides the ability to enable richer context and relationship-mapping to intelligence reports, electronic messages, and other unstructured narratives that are critical to counter-terrorism efforts.
- National security organizations and coalition partners, whether across large regions or communities of interest, stand to benefit greatly from the ability to not only load data as-is, but also to securely share and analyze the data in a timely fashion, by leveraging the massive scale of MarkLogic with Hadoop on Intel Architecture, as well as MarkLogic's advanced indexing capabilities.

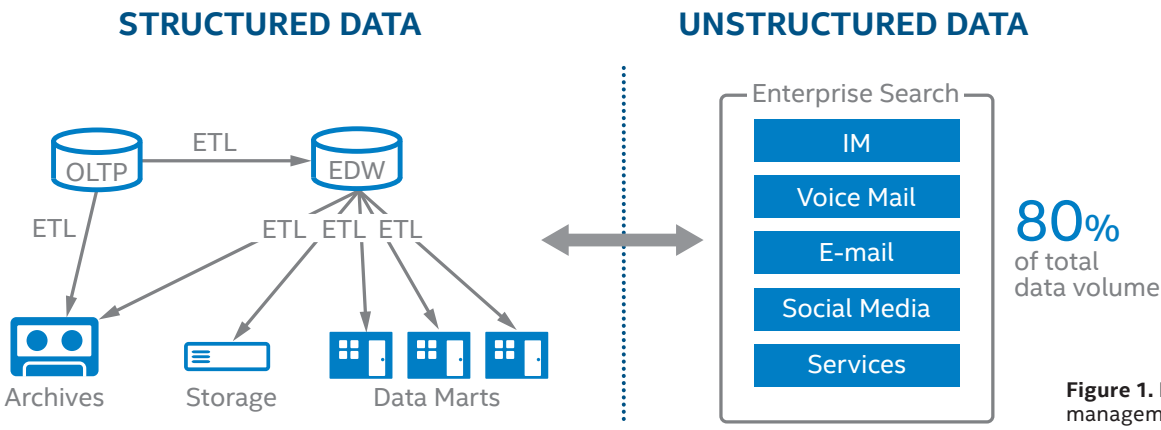

Figure 1. Last-generation data management architecture

The Criticality of Today's Data Governance

What drives mission success is actionable knowledge from data. The extent and value of that data—from daily transactions to emails, text documents, social media content, and others—is more critical today than ever. And, as agencies look to better leverage and share information both within their organization and with partners and the public, the policies and processes used to capture, manage, and protect that data have become even more important. These impacts are driving enterprises to take a new look at how they deal with data.

The Rigidity of Traditional Enterprise Data Environments

Organizations have long used online transaction processing (OLTP) systems based on Relational Database Management Systems (RDBMS) plus Storage Area Network (SAN) to gather the essence of their daily activities, which analytical systems periodically process. Organizational growth and queries beyond what the schema of the RDBMS was designed to provide, however, eventually result in a system that no longer serves the wide-ranging needs of the organization. These effects potentially cause lost revenue, missed opportunities, and more.

To adapt to information demands within the enterprise, IT often spins off enterprise data warehouses (EDW) from the company's RDBMS to create dedicated report and analytical systems that serve a specific application (Figure 1). These siloed data environments require more investment, more Extract-Transfer-Load (ETL) operations, and duplicated data, creating greater burden for IT, rising costs, and increased risk.

Eventually, more creative or complex analyses are required that the EDW cannot provide. Thus, individual departments within the organization create smaller data marts extracted from the EDW. These data marts provide the core content for real-time analysis using Excel* and other business intelligence tools. The results are even more copies of data pools and information systems potentially beyond IT's visibility and manageability.

At some point, scaling up the database, EDW, data marts, and storage for these proprietary systems becomes economically unsustainable. In order to maintain predictability in cost and performance of the most important data, enterprises archive the less important (usually older) data. But, finding the right slice of data across a large RDBMS schema is challenging. Again, it introduces brittle and costly ETL, and the archives are unavailable for deep analytics that might be needed. Additionally, the data marts unknown to IT might lag, working with older and possibly inaccurate information.

Unstructured Data Drives Change

This evolution has been repeated across enterprises in order to fulfill new mission requirements. It challenges any organization's data governance capabilities. And, it does not even include schema-less or "unstructured" data.

All of the text documents, sensor data, email, social media content, and video are simply not captured. Yet, getting a holistic view of a situation, such as a cybersecurity breach or other crisis, requires that unstructured data be part of the enterprise-wide data management.

Big Data solutions – including NoSQL data platforms – have arisen to help address challenges like these, in addition to providing new capabilities to gain insight from more varieties of data.

MarkLogic Enterprise NoSQL

For more than a decade, MarkLogic has delivered a powerful, agile, and trusted enterprise-grade, schema-agnostic Not-Only SQL (NoSQL) database platform. Using MarkLogic Server, an organization's entire structured and unstructured data repository can be stored in one indexed location, enabling fast application building on top of it and eliminating the need for siloed systems. MarkLogic can also be used in a data virtualization/Logical Data Warehouse environment to bridge data silos that cannot be merged in one location for whatever reason. This approach allows organizations to more quickly turn all data into valuable and actionable information in real-time, while reducing risk, cost, and management overhead.

Among its key features, the MarkLogic platform includes:

- Atomicity/Consistency/Isolation/Durability (ACID) transactions
- Horizontal scaling and elasticity
- Real-time indexing
- Full-text search and query
- High availability
- Disaster recovery
- Replication
- Government-grade security

Enterprises around the world and across a wide range of industries, including healthcare, entertainment, financial services, retail, government agencies, and others, have adopted the MarkLogic platform to manage and analyze all their data. These organizations are benefiting from building value from their data instead of schemas and infrastructure to support and understand it.

More recently, the emergence of Apache Hadoop* has brought yet more capabilities for enterprise data storage and analysis. MarkLogic integrates readily with Hadoop in a number of important ways.

The Era of Big Data and Apache Hadoop*

To take advantage of the value of all their data, organizations are aggressively moving toward storing and maintaining data that might have been previously discarded. Saving this legacy information creates a sandbox for data science, enabling possibilities of new and deeper population-level analyses, as well as more traditional data preparation and aggregation (ETL). But, how does an enterprise operationalize this rich information?

"For more than a decade, MarkLogic has delivered a powerful, agile, and trusted enterprise-grade, schema-agnostic Not-Only SQL (NoSQL) database platform."

As we have seen, the RDBMS model, while still offering the enterprise capabilities organizations have come to expect, constrains what can be done with the data. Mission success requires a new paradigm.

Apache Hadoop has emerged as a cost-effective place to store raw, intermediate, and finished data of all types—both structured and unstructured (Figure 2). It can accommodate massive amounts of data in any shape—and do it cheaply.

Hadoop also integrates tools for distributed computation across petabyte-sized volumes that are beyond what RDBMS + SAN implementations can do. Apache Hadoop has become the core of Big Data solutions with its staging, persistence, and analytics capabilities:

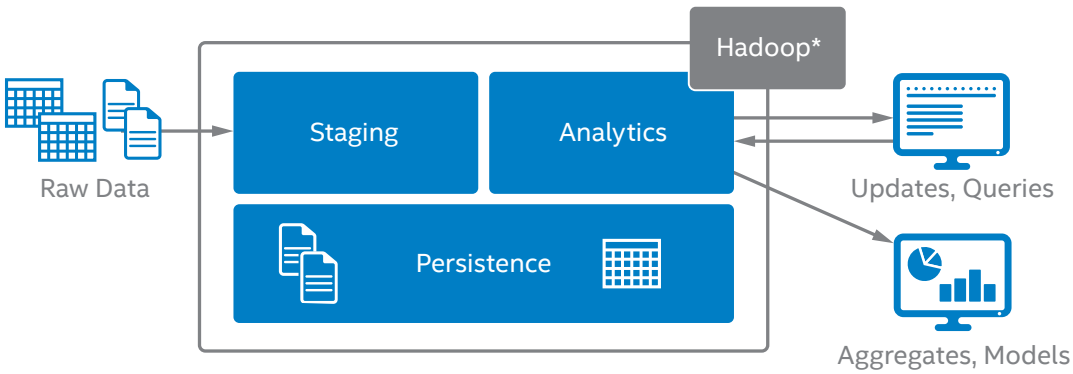

Figure 2. The capabilities of Apache Hadoop*

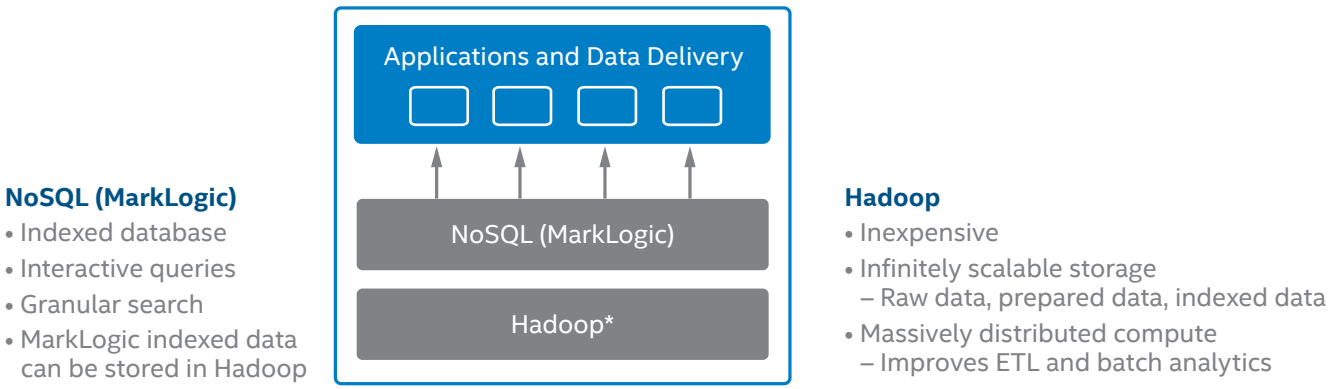

Figure 3. Enterprise NoSQL + Apache Hadoop*: new generation

- **Staging:** Load raw data into Hadoop. Use MapReduce* operations to prepare data for other uses, including filtering, aggregation, mash-up, transformation, etc.
- **Persistence:** Keep the raw inputs around for later inspirational integration and analytics, without losing the original context. Keep the intermediate prepared data around, also. Manage raw and prepared indexes under the same infrastructure and with the same governance policies.
- **Analytics:** Perform large-scale, population-level analyses on raw or prepared data.

However, while open source Apache Hadoop offers analytics and storage capabilities businesses want today, it was not designed for the real-time applications or the data governance requirements enterprises need.

MarkLogic with Apache Hadoop

To integrate with Hadoop, MarkLogic enhanced the MarkLogic platform to utilize Hadoop Distributed File System (HDFS) storage (Figure 3). Enterprises can run the MarkLogic database on top of HDFS, providing role-based security, full-text search, ACID transactions, and the flexibility of a granular document data model for real-time applications—all within the existing Hadoop infrastructure.

With MarkLogic's data files stored in HDFS, analysts can also run MapReduce jobs on those files directly. This opens up MarkLogic's formerly proprietary data format to other workloads and makes the file format a viable long-term archive option.

With MarkLogic's indexes stored in HDFS, companies can quickly gather ad hoc subsets of indexed data and attach them to a MarkLogic database to have that data immediately available for interactive updates and queries. This simplifies operations and data governance, maintains the security and metadata when it was first indexed, and allows use of those initial indexes (and security and metadata) throughout the life of the data.

Apache Hadoop

Proven in production at some of the most demanding enterprise deployments in the world, Apache Hadoop is supported by experts at Intel with deep optimization experience in the Apache Hadoop software stack as well as knowledge of the underlying processor, storage, and networking components.

Intel's enhancements to Hadoop are designed to enable the widest range of use cases on Hadoop by delivering the performance and security that enterprises demand. Intel delivers platform innovation in open source, and it is committed to supporting the Apache developer community with code and collaboration.

The combination of MarkLogic, Intel technologies, and Apache Hadoop optimized for Intel Architecture delivers the best of MarkLogic's platform performance, security, and manageability. While a NoSQL + Hadoop solution helps bridge traditional RDBMS with the wider-ranging data analytics and storage capabilities of Hadoop, the combination of MarkLogic and Apache Hadoop enhanced on Intel Architecture makes this an enterprise-class Big Data solution.

MarkLogic and Intel

The combination of MarkLogic and Apache Hadoop enables enterprises to implement both granular, real-time analyses plus deep batch analytics on massive data sets with enterprise-grade data governance—all on top of a single repository (Figure 4). With MarkLogic and Apache Hadoop, rather than building a new dedicated silo of storage, database, warehouse, middleware, and thick client, organizations can focus on the value of the data instead of the infrastructure and still be assured there is no compromise on performance, availability, or security.

MarkLogic provides the secure, reliable, and high-performance real-time indexing, search, and analysis platform the company's customers have come to trust. Enhancements based on Intel® technologies offer hardware-enhanced, secure Apache Hadoop operations with significant performance improvements over pure open source Hadoop.

The combination of MarkLogic and Intel dramatically shrinks the application stack, making building applications much less expensive and less risky. Organizations can more freely innovate and cut their losses early if an idea doesn't work.

"The combination of MarkLogic and Apache Hadoop for Intel Architecture delivers the best of MarkLogic's new-generation platform and Intel's hardware-enhanced Apache Hadoop performance, security, and manageability."

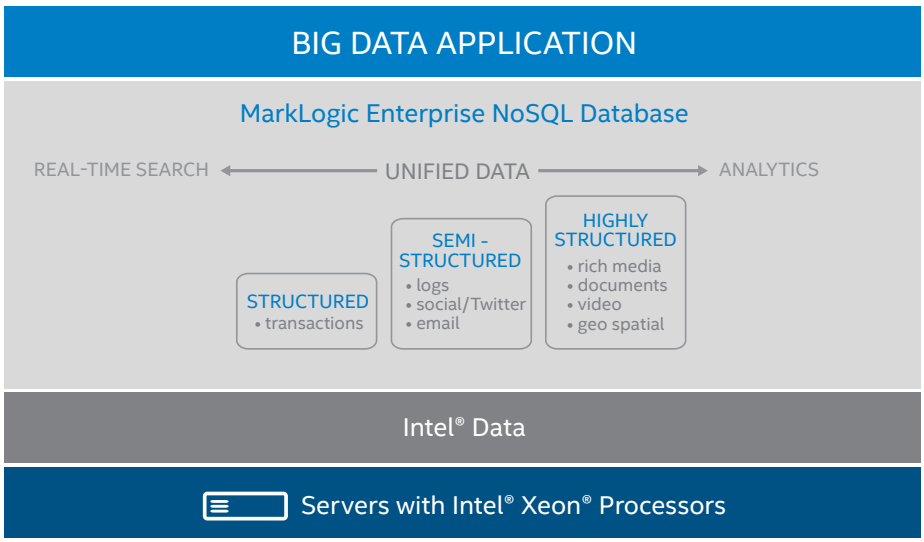

Figure 4. MarkLogic and Apache Hadoop* on Intel® technology

BBD—Before Big Data

Only in the last few years have unstructured data, Big Data, and Apache Hadoop* become important parts of enterprise operations. But, before the 2009 release of Apache Hadoop, in 2003 MarkLogic released its Not-Only SQL (NoSQL) database, search engine, and application software platform to enable analytics on both structured and unstructured data (Figure 5). The capability of a data storage, query, and analysis platform beyond Relational Database Management Systems (RDBMS) was born—long before the idea of Big Data. This new-generation MarkLogic platform became the foundation of systems that have given organizations the ability to gain insight from and act on more of their data in new ways.

With the emergence and growing adoption of Hadoop across industries and the Big Data storage and processing benefits it offers, MarkLogic integrated Apache Hadoop into the MarkLogic platform (NoSQL + Hadoop). The combination gives Hadoop the real-time search capabilities and enterprise-grade database platform organizations need to operationalize all their data, yet Hadoop still requires additional critical capabilities, like encryption and management tools, IT demands.

MarkLogic has integrated Intel into their system to deliver these needed features—not just NoSQL + Hadoop, but MarkLogic + Intel.

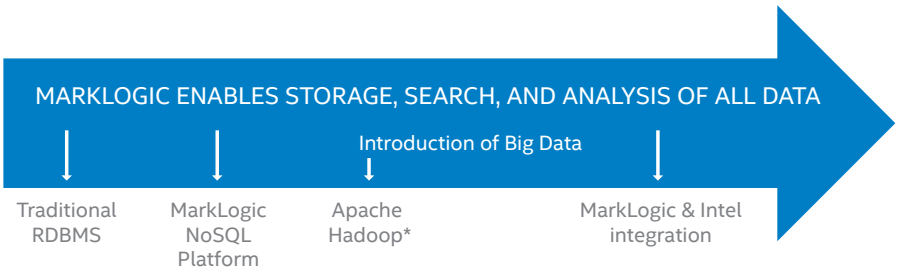

Figure 5. MarkLogic enables 'Big Data' before Big Data

Summary

Government agencies are the original Big Data creators and users, and they have stringent requirements for data management – including extensive security, auditing, and retention policies. By combining MarkLogic with Hadoop and Intel Architecture, government organizations can cost-effectively integrate, store, search, share, and archive their data as needed to ensure mission success.The integration of MarkLogic and Hadoop creates a platform that helps IT reduce risk and contain—even reduce—cost, while accelerating application development of data analytics to serve the needs of an public agencies.

MarkLogic and Intel-enhanced Apache Hadoop allow government organizations to keep all their data readily available for business intelligence and deeper population-level studies that can provide new insights and reveal new opportunities.

For more information on MarkLogic and Intel, visit www.marklogic.com and www.intel.com.

MarkLogic on Intel® Technology

Intel® technology, as a hardware foundation for Apache Hadoop, delivers significant performance improvement for Hadoop processing (Figure 6). Along with Intel® Xeon® processors, Intel® Solid State Drives, and Intel® 10GbE networking, Intel offers a robust contribution to support the new generation of the MarkLogic Enterprise NoSQL platform.

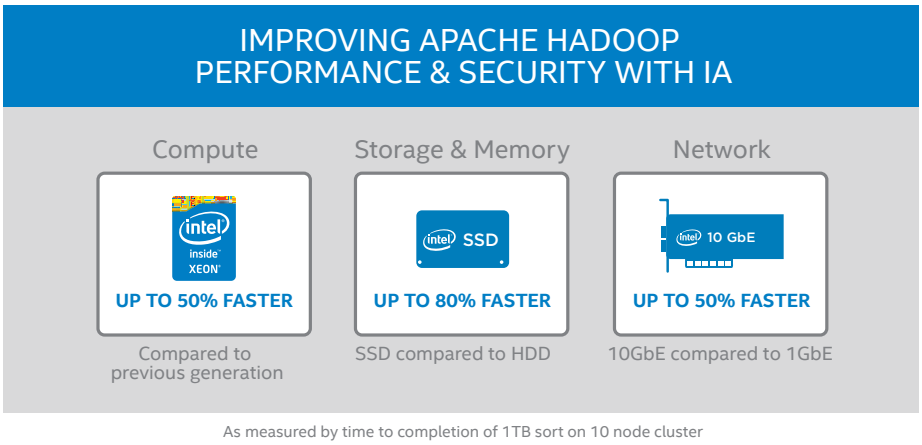

Figure 6. Intel® technology provides accelerated performance for Apache Hadoop*

[1] **Hadoop Westmere Test Bed: 4 hours**

Hardware Configuration: Arista 7050T; 10 x SuperMicro 1U servers: Intel Processor: 2 x 3.46 GHz Intel® Xeon® processor 5690; Memory: 48 GB RAM; Storage: 5 x 700 GB 7200 RPM SATA disks; Intel® Ethernet 10 Gigabit Server Adapters (10GBASE-T); Intel® Ethernet Gigabit Server Adapter (1000BASE-T)

Software Configuration: Operating System: CentOS 6.2; Hadoop: Cloudera's Distribution; Java*: Oracle JDK 1.7.0.

Cluster Configuration: 1 Client machine; 1 Head node (Name node, Job Tracker); 9 Workers (data nodes, task trackers).

Network Division Hadoop Romley Test Bed: 7 minutes

Cluster Configuration: 1 Head Node (name node, job tracker); 10 Workers (data nodes, task trackers); 10-Gigabit Switch: Cisco Nexus 5020;

Software Configuration: Intel Distribution for Hadoop 2.1.1; Apache Hadoop 1.0.3; RHEL 6.3; Oracle Java 1.7.0_05.

Head Node Hardware: 1 x Dell r710 1U servers: Intel: 2x3.47GHz Intel® Xeon® processor X5690; Memory: 48 GB RAM; Storage: 10K SAS HDD; Intel® Ethernet 10 Gigabit SFP+; Intel® Ethernet 1 Gigabit.

Worker Node Hardware: 10 x Dell r720 2U servers: Intel: 2 x 2.90 GHz Intel® Xeon® processor E5-2690; Memory: 128 GB RAM; Storage: 520 Series SSDs x 5; Intel® Ethernet 10 Gigabit SFP+; Intel® Ethernet 1 Gigabit.

 Printed in USA 0614/JO/OCG/PDF ♻ Please Recycle 330577-002US