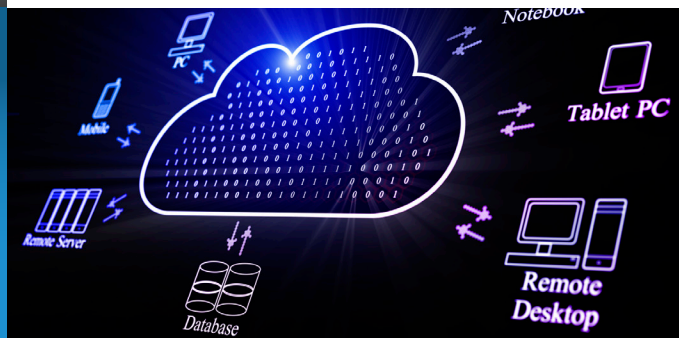
MarkLogic + Intel:
Partnering to Extend Hadoop to Enterprise-Class NoSQL Databases



The combination of MarkLogic® NoSQL database technology and Intel® Distribution for Apache™ Hadoop® software enables enterprises to implement real-time Big Data applications that deliver immediate business insights.

A Powerful Big Data Partnership

MarkLogic and Intel have joined forces to provide enterprises with a powerful package that combines the Big Data advantages of Apache™ Hadoop® and an enterprise-class NoSQL (not only SQL) database.

Since its first release in 2007, Hadoop has become the de facto standard for processing Big Data at multi-petabyte scale. With Intel Xeon® processors powering four out of five servers in data centers, Intel® Distribution for Apache Hadoop* software is the emerging distribution of choice for enterprises that want to implement open-source Hadoop technology, while also seamlessly exploiting the full power of hardware-level security and performance acceleration on their industry-standard hardware.

MarkLogic 6 Enterprise NoSQL Database is designed to deal with huge volumes of highly diverse data, allowing users to interact with all the data they need—right now, in real time. The market-leading MarkLogic NoSQL technology provides a reliable, scalable, and secure Big Data platform that delivers value by making more kinds of data immediately available, and also leveraging the enterprise's existing investment in data tools and expertise. Over the past decade, MarkLogic has deployed more customer-facing, mission-critical applications in large enterprises than have all other NoSQL database vendors combined.

Major Challenges

Business managers want an enterprise-proven, low-TCO solution that enables them to optimize the business intelligence they derive from their data, improve productivity and efficiency, and maintain a competitive advantage. IT managers require a common infrastructure that is scalable, flexible, cost-effective, and fault-tolerant, that supports many different applications and data types, that overcomes silo limitations, and that augments their current infrastructure rather than replacing it. And users need to have interactive, real-time conversations with a large store of diverse data.

Typical Data Needs

- Content authoring and delivery
- Data unification
- Data virtualization
- Digital asset management
- Metadata cataloging
- Open source intelligence
- Search and discovery
- Social media analysis
- Logical data warehouse

Key Benefits

- Manages all types of data, as-is, however it is structured
- Supports a spectrum of Big Data applications on one infrastructure
- Provides users with immediate, real-time data access
- Enables offline batch analytics where real-time is not necessary
- Leverages existing business-intelligence tools and IT expertise
- Scales easily and non-linearly as the data grows
- Simplifies integration of tools and customization of resources
- Ensures enterprise-class availability and offers granular security
- Runs on cost-effective, clustered, industry-standard servers
- Reduces development costs and speeds software evolution
- Delivers world-class MarkLogic professional support

Intel's Contribution To Hadoop

Intel is committed to the wide adoption and use of Big Data technologies such as Hadoop. Complex data that requires compute-intensive analysis needs a combination of hardware and software management optimizations to deliver scale with a high return on investment. In addition to being a major contributor to the Hadoop open-source initiative and devoting

Hadoop in Brief

Apache Hadoop is an open-source framework for running applications on large server clusters using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

Hadoop implements a computational paradigm called MapReduce that divides the application into small fragments of work, each of which may be executed or re-executed on any node in the cluster. Hadoop Distributed File System (HDFS) stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. Both MapReduce and HDFS are designed to automatically detect and handle failures at the application layer. Hadoop is supported and continually enhanced by the Apache open-source software community.

Hadoop offers a computing solution that is scalable, flexible, fault tolerant, and cost effective. New nodes can be added without changing how data is formatted, how the data is loaded, and how jobs are written. Since Hadoop is schema-less, it can deal with any type of data, structured or unstructured, from any source. Data from multiple sources can be joined and aggregated to allow deeper analyses. If a node goes down, the system redirects work to another location and continues processing. And by running on industry standard servers, Hadoop reduces the cost per terabyte of storage.

substantial resources to Hadoop analysis, testing, and performance characteristics, Intel is leading the way to improve security, enable hardware-based acceleration technologies, and streamline the tasks of system management through the Intel Manager.

Intel Distribution for Apache Hadoop* software delivers:

- Security without comprising performance using hardware-assisted encryption (AES-NI)
- Performance boosting through advanced hardware features such as Streaming SIMD Extensions (Intel SSE) and Cache Acceleration Software (Intel CAS), as well as core Xeon, Solid-State Drive (Intel SSD), and 10-gigabit Ethernet capability optimizations
- Ease of management through auto-tuning of Hadoop configurations using machine-learning algorithms with Intel Manager

With growing volumes of unstructured and semi-structured data flooding into data centers, enterprises are finding that traditional relational databases are too limiting and inflexible. For more than a decade, MarkLogic has delivered a powerful and trusted enterprise-grade NoSQL database that enables organizations to unify their data and turn it into valuable and actionable information.

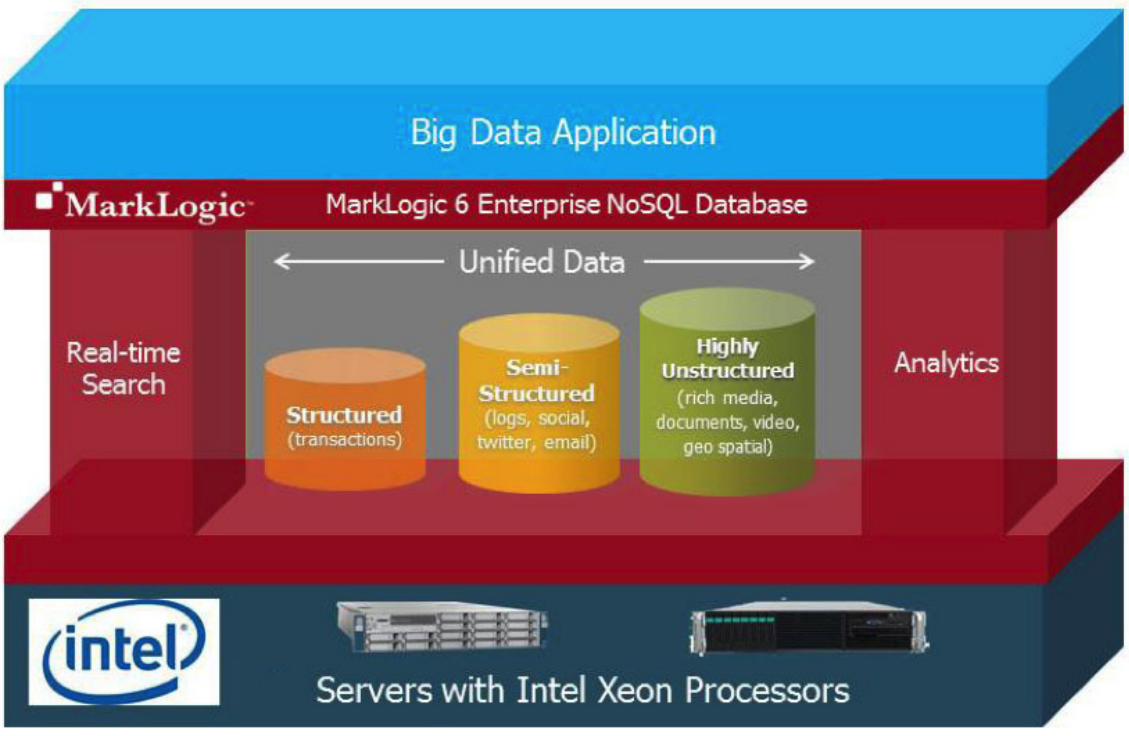


Figure 1: The architecture of MarkLogic's unique Enterprise NoSQL database

The MarkLogic team has developed a version of MarkLogic Server featuring HDFS storage. This distributed file system provides a real-time database for Hadoop that features scalability, performance, and availability, enabling a fluid mix of data between operational and analytic workloads. Enterprises can run the MarkLogic database on top of HDFS to provide atomicity/consistency/isolation/durability (ACID) transactions, role-based security, full-text search, and the flexibility of a granular document data model for real-time applications, all within the existing Hadoop infrastructure

Advantages for Hadoop Users

Get the best of Big Data applications by combining MarkLogic with Intel Distribution for Apache Hadoop software.

Real-Time Hadoop Applications

With MarkLogic running on HDFS, Hadoop can be extended to support low-latency applications such as real-time search and analytics. Using HDFS as storage for a NoSQL database is not new, but MarkLogic is the first vendor to support transactional updates, full-text search and alerting, a document data model, and enterprise security together in one integrated system.

Leverage a Common Infrastructure

Using MarkLogic and HDFS enables common batch-processing infrastructure to be used across many different projects and applications.

Unstructured Data Throughout

MarkLogic and Intel Distribution for Apache Hadoop software are both built to deal with unstructured data. By combining the two, the enterprise can eliminate the need to map data to rows and columns for every type of analysis or service endpoint. The MarkLogic/Hadoop combination makes it much easier to integrate new and unanticipated data sources. And it also allows managers to solve problems that would be difficult or impossible to address with either technology alone.

Optimize Hadoop for ETL

MarkLogic provides a single infrastructure for extracting data from outside sources, transforming it to fit operational needs, and loading it into storage—a process known as ETL. Raw data stored in HDFS may be refined and transformed by Hadoop before being consumed by MarkLogic Connector for Hadoop, a drop-in extension for Hadoop that provides efficient two-way communication between MarkLogic and HDFS using standard MapReduce jobs. This connector simplifies parallel loading from HDFS to MarkLogic and provides the ability to leverage MarkLogic's indexes for MapReduce processing (see Figure 2).

Overcome Silo Limitations

Organizations with business-critical information scattered across separately owned and managed silos need to put this data in one place. MarkLogic's Enterprise NoSQL database overcomes the limitations of rigid RDBMS-based systems, enabling comprehensive data virtualization. And using MarkLogic and Hadoop HDFS enables a common batch-processing infrastructure to be used across many different projects and applications.

Hadoop Archival Storage

Organizations need ready access to operational data, along with the ability to retain older information in less expensive archival storage, where it can be accessed as needed. Enterprises can keep their infrequently-accessed, non-operational data in HDFS/Hadoop and load it into MarkLogic when required.

Enhanced Security

MapReduce running in situ on data in the MarkLogic database can leverage MarkLogic's granular security model to restrict analytic jobs to only the data they are entitled to, which is critical for many government and enterprise applications.

Tools and Support

In ongoing efforts to more closely align MarkLogic and Hadoop, MarkLogic redistributes the Hortonworks Data Platform together with a suite of tools that provides connectivity between Hadoop and MarkLogic. MarkLogic also provides enterprise-class support for mission-critical applications that combine MarkLogic and Hadoop.

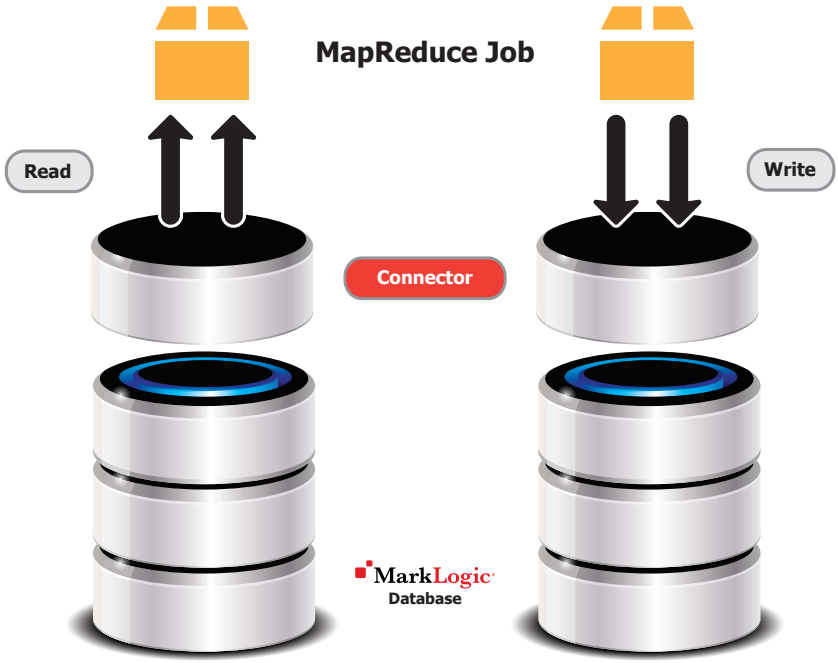


Figure 2. MarkLogic Connector for Hadoop

MarkLogic: A Next-Generation Database

Data analysts today must deal with growing volumes of information that make it difficult and time-consuming to separate useful and actionable information from the chaff. They risk drawing incorrect conclusions because they overlook buried information. Or they miss a critical deadline or opportunity because they cannot get relevant data fast enough—or get it at all.

Next-generation Big Data requires a next-generation database. The MarkLogic Enterprise NoSQL database meets the needs of enterprise Big Data implementations because it is engineered for volume, velocity, variety, and complexity.

Volume

Volume refers to the massive quantities of data that organizations must harness if they are to improve decision making. Data volumes are increasing at an unprecedented rate. The MarkLogic database is optimized for today's advanced hardware. It supports fast look-up and a "shared-nothing" architecture that features independent nodes and no contention.

Velocity

The speed at which data is created, processed, and analyzed continues to accelerate. Contributing to higher velocity is the real-time nature of data creation, and the need to incorporate streaming data into business processes. Velocity impacts latency, the lag time between when data is created or captured, and when it is accessible. For time-sensitive processes, certain types of data must be analyzed in real time to be of value. MarkLogic's database is both fast and agile. It delivers data in real time with performance alerting, and provides multiversion concurrency control so that writing and reading transactions do not block each other.

Variety

Organizations need to integrate and analyze data from an array of both traditional and non-traditional information sources, from within and outside the enterprise. Data can be generated in countless forms, including: text, web, tweets, audio, video, click streams, log files, application-specific documents, and other formats. MarkLogic can handle all these types of data, with metadata being extracted and stored to aid in search.

Complexity

MarkLogic leverages XML to handle variable-length elements, hierarchical relationships, sparse data, and schema-independent data. Information that should not or cannot be disassembled into rows and columns—such as contracts, manuals, books, emails, tweets, and metadata—is ideally suited to the XML-based, document-centric model used in MarkLogic. Universal indexes allow loading information as-is, thus avoiding rigid, predefined schemas. This is especially efficient for indexing and querying information with poorly defined, poorly followed, changing, or unknowable schemas. MarkLogic also readily supports information that adheres to a schema, and can even enforce a schema if that is what an organization needs.

For any enterprise looking to enhance its business operations by building Big Data applications using Hadoop, the MarkLogic + Intel partnership offers a powerful, proven, well-supported solution. MarkLogic 6 Enterprise NoSQL Database manages all types of data, at scale, in real time, providing a reliable, scalable, and secure Big Data platform that delivers value by leveraging existing tools and expertise. Intel Distribution for Apache Hadoop software delivers higher performance, scalability, ease of management, and security for Big Data analytics at lower cost.

More Information

For more information about Intel Distribution for Apache Hadoop software, visit www.intel.com/bigdata.

For more information about MarkLogic, visit www.marklogic.com or contact:

David Ponzini
SVP, Corporate Development / VP, Asia Pacific
david.ponzini@marklogic.com
650.655.2328 (Office)
925.997.9872 (Mobile)

Jeff Faraday
Director, Alliances
jeff.faraday@marklogic.com
650.655.2372 (Office)
925.872.6545 (Mobile)

MarkLogic Corporation
www.marklogic.com
sales@marklogic.com
+1 877 992 8885

Headquarters
999 Skyway Road, Suite 200
San Carlos, CA 94070
+1 650 655 2300