MarkLogic®

DATASHEET

Hadoop Integration

Organizations need to store and analyze massive amounts of structured and unstructured data from disparate data sources – more data than ever before with the shortest possible time to insight so organizations can act at the speed of their business. Hadoop* is a great tool that has been adopted by a large number of organizations to help with this task. Marklogic® provides the most comprehensive set of complimentary features for integration with Hadoop.

Hadoop: HDFS and MapReduce

Hadoop has become popular because it is designed to cheaply store massive amounts of data in the Hadoop Distributed File System (HDFS) and run large-scale MapReduce jobs for batch analysis.

- **HDFS** is a Java-based file system that provides scalable and reliable data storage across clusters of commodity servers. In production, HDFS has been shown to scale to 200 Petabytes of storage across 4,500 servers, supporting close to a billion files.
- **MapReduce** is a processing framework that uses a "divide-and-conquer" paradigm that takes a huge task and breaks it into small parts ("Map") and then aggregates the resulting outputs from each part ("Reduce"). Any large task that can be broken into smaller pieces is a candidate for use with Hadoop.

MarkLogic: The Best Database for Hadoop

Hadoop is great for storing and analyzing data, but it still needs a database. Hadoop is simply not designed for low latency transactions in real-time interactive applications, or applications that require enterprise features such as government-grade security and high availability and disaster recovery. Even as the Hadoop ecosystem continues to evolve with new components such as Spark, the real benefits of Hadoop are realized only when running alongside an enterprise-grade database.

MarkLogic is the best database for Hadoop. It can seamlessly run alongside the Hadoop ecosystem, connecting via the **MarkLogic Connector for Hadoop** to act as the database to power real-time, transactional applications. Additionally, you can also just use HDFS as storage, without even needing the Connector. MarkLogic can also leverage HDFS using Tiered Storage, seamlessly moving data between any combination of HDFS, S3, SSD, SAN, NAS, or local disk to support specific SLAs and cost objectives without modifying downstream application code.

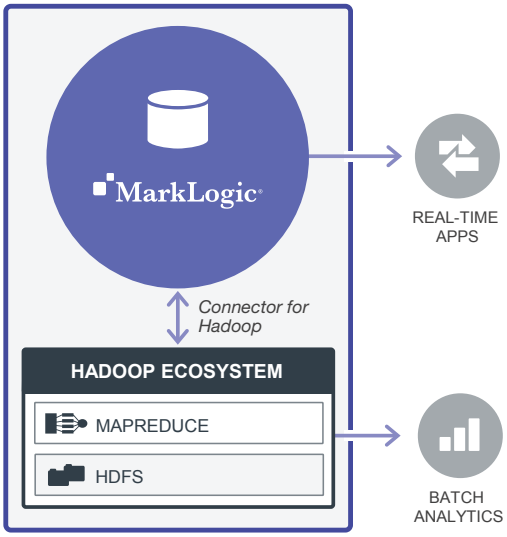
MARKLOGIC	HADOOP
Online, low latency applications	Offline, high latency processing
Real-time transactions	Long-haul, batch analytics
Built-in search	Distributed, cost-effective storage

MarkLogic®

Modern Hadoop Infrastructure

MarkLogic stores data in distributed clusters much like Hadoop, an architectural parity that makes it easy to move data partitions (called forests) between MarkLogic hosts and the Hadoop ecosystem. This is done with the **MarkLogic Connector for Hadoop**, using MarkLogic as either an input source or output destination. The benefit of using MarkLogic is that it can act as a database for data integration and to power real-time transactional applications.

Current versions of Hadoop all offer secure HDFS, ensuring access to the data is authenticated using a trusted protocol like Kerberos. As the world's most secure NoSQL database and as the only NoSQL database with Common Criteria Certification, MarkLogic complements and extends Hadoop's security model.



Popular Use Cases

- **Large ETL Processes** – Use Hadoop to process raw data, particularly for resource intensive processes using highly specialized libraries such as facial recognition in images, machine learning, or sophisticated entity extraction. Then, send the output to MarkLogic for *ad hoc* queries using MarkLogic indexes.
- **Archival Storage** – Manage data throughout its lifecycle by using MarkLogic to store high value, operational data and HDFS as a storage tier to archive older data that has less probability of re-use. Data in HDFS can still be brought online in MarkLogic at a moment's notice, giving you the ability to meet operational objectives at a lower cost.
- **Securing Data in Hadoop** – MarkLogic is the most secure NoSQL database. Its security model can be applied to data stored in HDFS, so analytic jobs can be restricted. This comes out-of-the-box, so you do not have to integrate other components such as Zookeeper and Accumulo.

About MarkLogic

MarkLogic is the world's best database for integrating data from silos, providing an operational and transactional Enterprise NoSQL database platform that integrates data better, faster, with less cost. Visit www.marklogic.com for more information.

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