TERRACOTTA®
GO BIG. GO FAST.

WHITE PAPER

BigMemory and Hadoop:
Powering the Real-time Intelligent Enterprise

BIGMEMORY: IN-MEMORY DATA MANAGEMENT FOR THE REAL-TIME ENTERPRISE

Terracotta BigMemory is the solution of choice for enterprises seeking the speed and performance advantages of keeping high-value big data in-memory. In industries ranging from financial and marketing services to media and travel, business innovators are using BigMemory as a central in-memory store for making terabytes of data available – in real-time – to the enterprise. With BigMemory, all the RAM in enterprise servers is available for immediate data access. Unlike other solutions, BigMemory delivers predictable, extremely low latency while slashing required application server footprints by 100x or more – even at terabyte scale.

Virtually every enterprise is now grappling with the challenge of managing big data. Increasingly, business innovators seek not only to bolster their ability to manage large data sets, but also to make those data sets instantly accessible for analysis. Insights gained from big data are a source of significant competitive advantage and enterprise profitability.

Big data analytics and Hadoop have practically become synonymous. In this paper we look at how Terracotta enterprise customers are taking a game-changing approach, adding BigMemory, the world's leading in-memory data management solution, to their Hadoop deployments to deliver on the promise of the real-time intelligent enterprise.

The Hadoop Challenge:

SHARING AND BUILDING NEW INTELLIGENCE

Enterprises use Hadoop[1] for everything from setting the optimal price of a plane ticket to figuring out which products pair best for co-marketing campaigns. Hadoop is an open source Java framework for performing computations, often using commodity hardware, over large data sets in a distributed network.

Hadoop is an implementation of Google's MapReduce computation paradigm[2] coupled with a distributed data storage mechanism called the Hadoop Distributed File System (HDFS). It can scale from one computation node to thousands, and it can operate on petabytes of data. Its built-in failure-handling capabilities ensure fault-tolerant execution of distributed computations.

Nevertheless, enterprises face two important challenges with Hadoop, both related to accessing data in real-time:

1. **Hadoop's datasets aren't as up-to-date as possible because data is typically batch-loaded from disk-bound databases.** Like any analytics platform, Hadoop can only draw insights from the data it has available. Batch-loading data from a disk-bound database into Hadoop degrades the currency of data and limits the enterprise's ability to identify and act on short-term trends.

2. **Hadoop's data resides in disk-bound architectures, which slows the velocity of insights delivered to the business.** Even though Hadoop may determine that customers with size 12 shoes are more likely to buy basketballs, that insight only becomes valuable when it's available to the enterprise's business systems. Stuck in Hadoop, the insight cannot be used to improve sales.

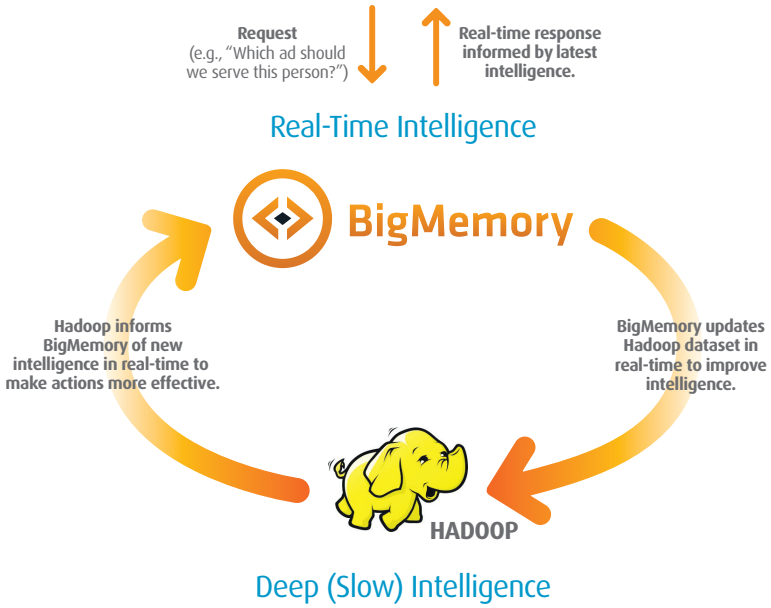
In short, Hadoop's batch-oriented architecture makes it challenging to get real-time activity into Hadoop and valuable insights out of Hadoop.

[1] http://hadoop.apache.org/
[2] http://research.google.com/archive/mapreduce.html

Hadoop and BigMemory:

A VIRTUOUS CYCLE FOR REAL-TIME INTELLIGENCE

To resolve these issues, enterprises use BigMemory as an in-memory store for immediately sharing Hadoop insights throughout their organizations and for feeding new activity into Hadoop in real-time. Working together, Hadoop and BigMemory create a virtuous cycle whereby real-time requests (e.g., "What ad should we serve to this person?") are answered based on fresh intelligence from Hadoop. Similarly, new transactions are immediately fed into Hadoop to improve its analytics.



As increasing numbers of transactions and other requests flow in, BigMemory enables applications to respond intelligently in real-time, based on insights from Hadoop. Meanwhile, Hadoop provides the most up-to-date answers because BigMemory adds transactions to Hadoop's data store as soon as they're available.

BIGMEMORY AND HADOOP: FINANCIAL SERVICES SUCCESS STORY

A Fortune 500 financial services firm with over $100 billion in annual transaction volume wants to use BigMemory with Hadoop to improve the speed and accuracy with which it detects fraudulent transactions. When it comes to fraud detection, speed and accuracy can drive significant increases in profitability, since the company's fraud losses can amount to hundreds of millions of dollars annually.

This customer uses BigMemory to manage over 4 TB of in-memory transaction data. BigMemory keeps a combination of events, at-risk accounts, fraud rules, long-term transactions and short-term transactions in memory – distributed across a server array – for fast online processing.

The company also stores its entire transaction history in a large Hadoop cluster that executes a long-running, iterative risk-scoring algorithm over this extensive dataset. Terracotta's BigMemory-Hadoop connector (explained below) can seamlessly move all new transactions from BigMemory into Hadoop, quickly incorporating this new data into risk scoring. At the same time, the connector can return updated risk scores to BigMemory for real-time fraud detection based on current intelligence.

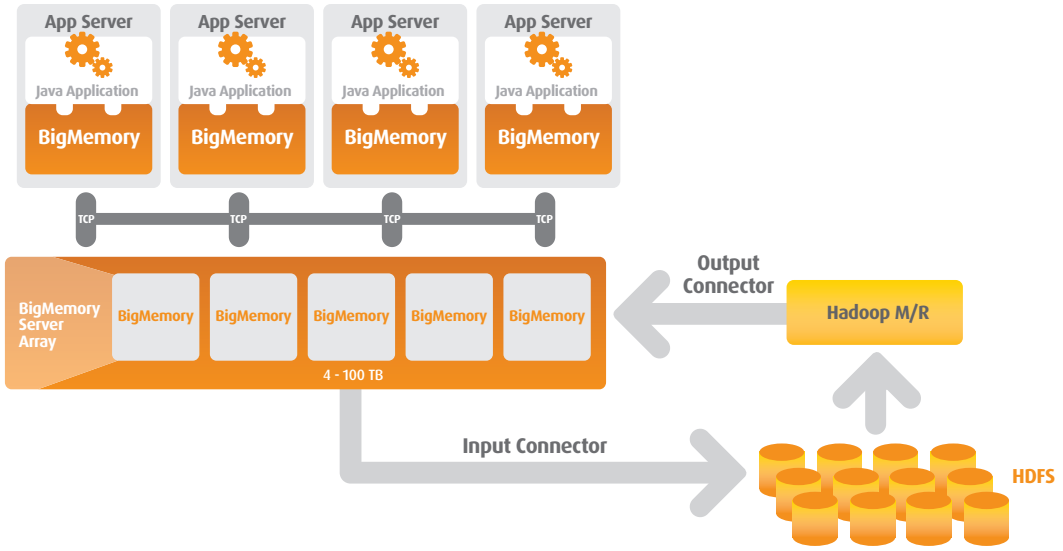


Figure 1: Fortune 500 financial services firm's architecture for BigMemory and Hadoop

Before BigMemory, the company failed to meet its service-level agreement (SLA) of 800 milliseconds per transaction around 10 percent of the time. Now, fraud detection consistently takes place in under 650 milliseconds. Fraud detection accuracy can be enhanced thanks to the real-time connection between BigMemory and Hadoop. It's estimated that BigMemory has saved the company tens of millions of dollars annually, and the initial deployment took only 8 weeks.

The Details:

BIGMEMORY-HADOOP CONNECTOR EXPLAINED

Plugging Hadoop into BigMemory is as simple as adding a JAR file and changing a line or two of code in the Hadoop application. The input and output connectors take care of automatically moving data back and forth between BigMemory and Hadoop. The BigMemory-Hadoop Connector eliminates the latency and complexity of batch jobs that might otherwise be used to get information into, and results out of, Hadoop.

The BigMemory-Hadoop Connector helps in three ways:

1. Increases the speed at which data from BigMemory is made available to Hadoop
2. Reduces the time it takes for Hadoop results to be accessible by online applications, from hours to seconds
3. Simplifies the movement of data between BigMemory and Hadoop

BIGMEMORY-HADOOP CONNECTOR: INPUT

Figure 2 illustrates the architecture of the BigMemory-Hadoop Connector's input flow. As the online application puts data elements into BigMemory, a write-behind service inside the BigMemory Server Array automatically stores the new data in Hadoop SequenceFiles in HDFS, giving Hadoop a real-time view.

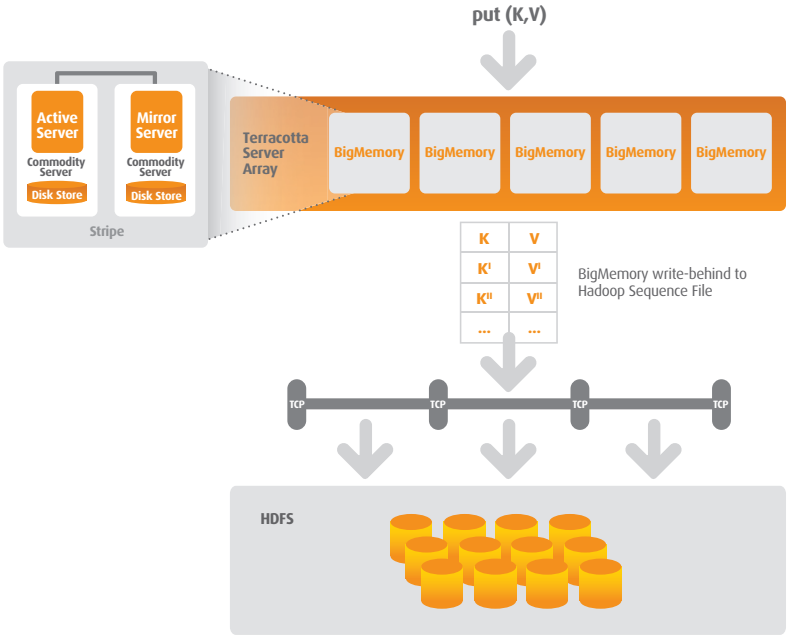


Figure 2: BigMemory-Hadoop Connector's input flow architecture

BIGMEMORY-HADOOP CONNECTOR: OUTPUT

Figure 3 shows the architecture of the BigMemory-Hadoop output flow. In a traditional Hadoop cluster, the results of a job are written to HDFS. Since HDFS files are write-once, they are locked for reading until the entire MapReduce job is completed.

The BigMemory-Hadoop Connector allows reducers to optionally write their results directly into BigMemory as results become available, making each key/value output pair visible immediately, rather than waiting for the entire job to complete and close the HDFS output file.

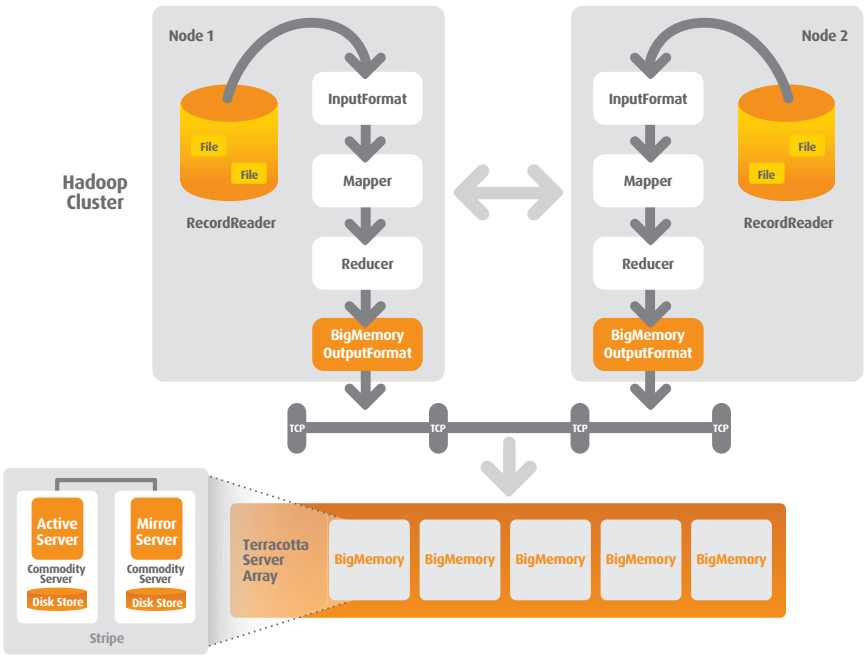


Figure 3: BigMemory-Hadoop Connector's output flow architecture

Want to learn more?

Terracotta software is already used in millions of deployments worldwide to make petabytes of data available in machine memory at microsecond speed. As big data goes mainstream, Hadoop is finding a home in many of those same IT shops. BigMemory's snap-in integration with Hadoop makes real-time intelligence powered by big data a reality for the enterprise.

FOR MORE INFORMATION:

Email sales@terracotta.org or download BigMemory for free at http://terracotta.org/bigmemory

About Terracotta, Inc.

Terracotta, Inc. is a leading provider of game changing big data management solutions for the enterprise. Its flagship BigMemory product line features big data in-memory solutions that deliver performance at any scale. Terracotta's other award winning data management solutions include Ehcache, the de facto caching standard and Quartz the de facto scheduler for enterprise Java. Terracotta supports the data management needs of a majority of the Global 1000 with over 500,000 deployments of its products. Terracotta is a wholly owned subsidiary of Software AG (Frankfurt TecDAX: SOW).

For more information, please visit www.terracotta.org.

TERRACOTTA, INC.

575 Florida St. Suite 100
San Francisco, CA 94110

For Product, Support, Training and Sales Information:
sales@terracotta.org

USA Toll Free
+1-888-30-TERRA

International
+1-415-738-4000

Terracotta China
china@terracottatech.com
+1-415-738-4088

TERRACOTTA®
GO BIG. *GO FAST.*

A SOFTWARE AG COMPANY