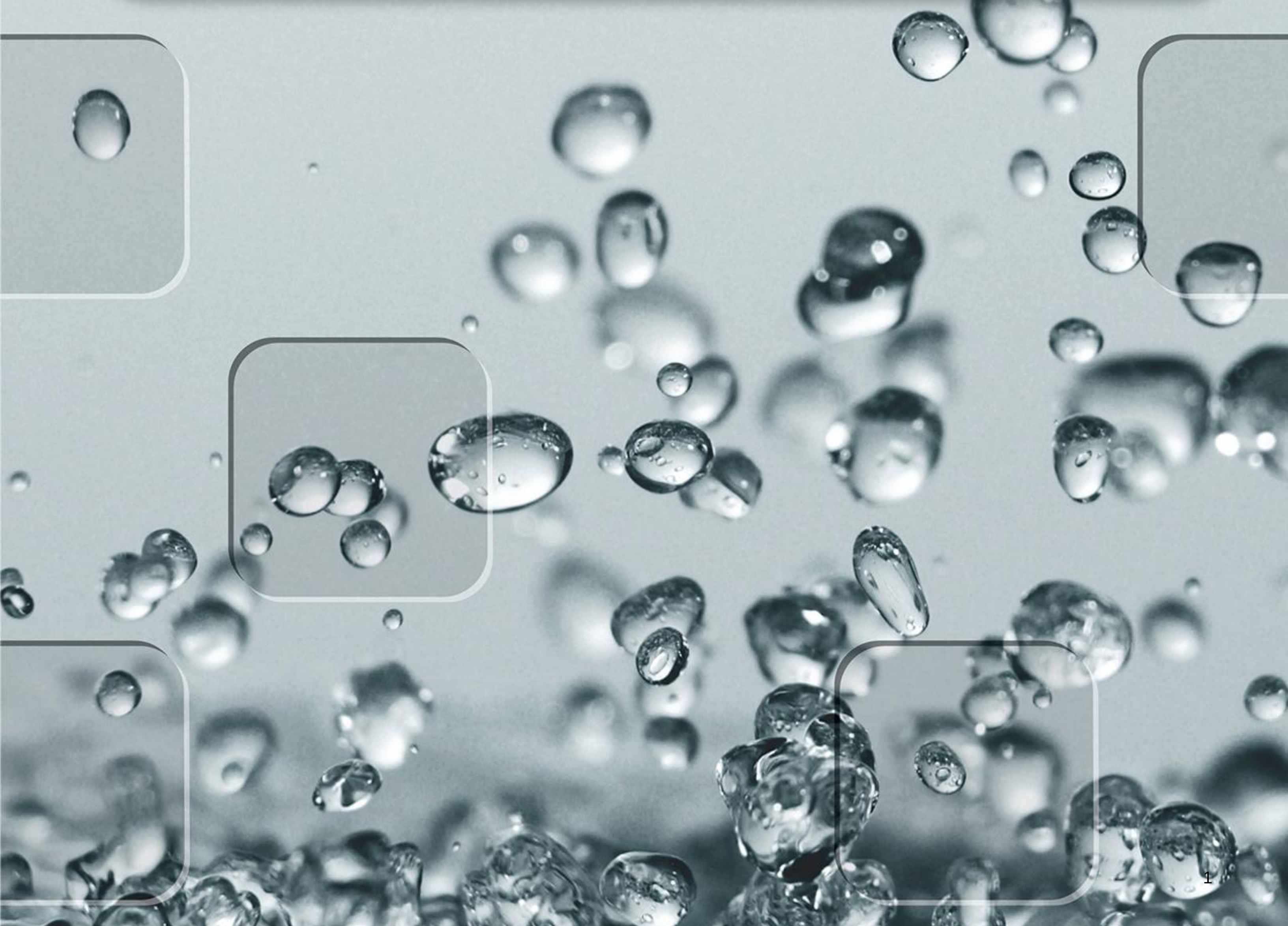
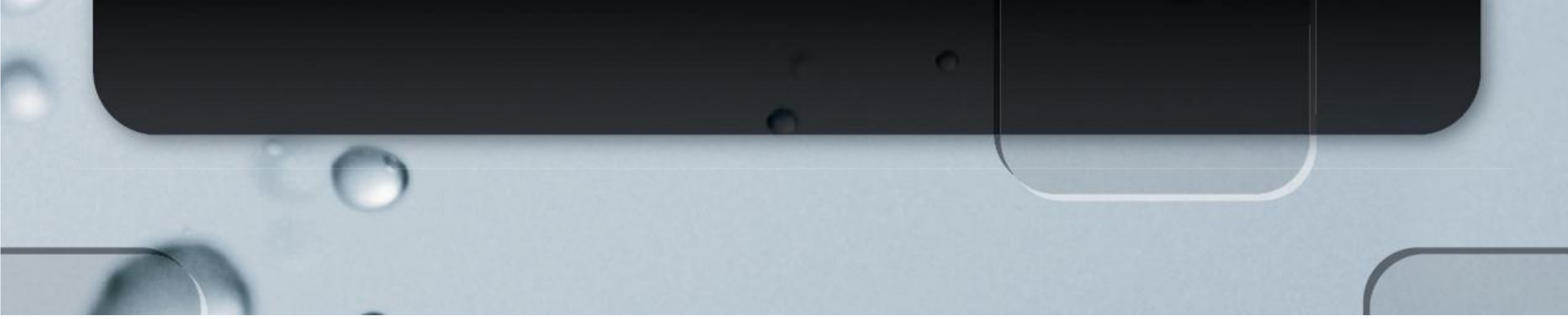
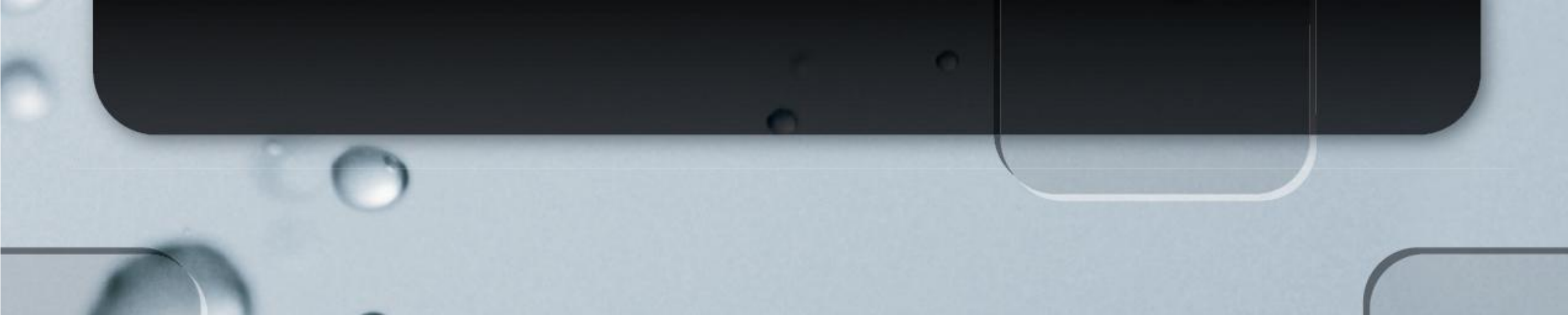
Vectorwise 3.0
Fast Answers from Hadoop

Technical white paper





Contents



Executive Overview

Vectorwise is the record-breaking analytic database for performance and price/performance (see Figure 1, page 3). Now with the Vectorwise Hadoop Connector, it is also possible to load data into Vectorwise from Apache Hadoop Distributed File System (HDFS) at extremely fast speeds on affordable commodity hardware.

A load throughput rate of over 3 TB an hour from HDFS into Vectorwise was achieved in our performance testing using a single 16 core Vectorwise server costing less than $15,000 USD, and 4 Hadoop nodes using identical hardware.

The Vectorwise Hadoop Connector makes it fast and easy to access data on Hadoop Distributed Files System (HDFS) and bring it into Vectorwise for fast answers. This paper will describe the use cases and performance testing results for the Vectorwise Hadoop Connector.

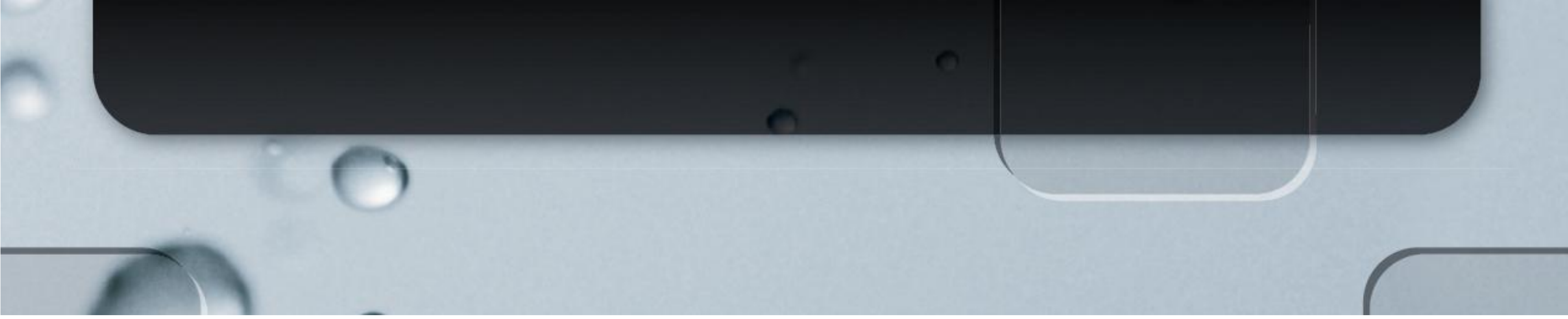
Introduction

A new breed of data-centric organizations are emerging. Leveraging fast, cost-effective Big Data technologies such as Actian Vectorwise and Hadoop, organizations are finding new ways to extract value from structured and unstructured data, and monetize Big Data.

Relational databases have long been the standard for managing and analyzing structured data. Meanwhile, new technologies have emerged to generate a wealth of less structured data types such as weblogs, social media, sensors, location and machine-generated data. The growth of these new data types have made relational databases a less effective and expensive storage engine.

Hadoop has become a popular Big Data framework because it provides near unlimited storage for unstructured or semi-structured data, as well as a massively parallel architecture for capturing and retrieving large volumes of content. However, Hadoop is a simple architecture with less built-in optimizations for highly interactive, multi-user query environments that most relational databases such as Vectorwise provide.

“Hadoop connectors” are necessary to make data on Hadoop available to the relational database for analysis. The Vectorwise Hadoop Connector makes it fast and efficient to load massive volumes of data from Hadoop into the fast and cost-effective Vectorwise analytic database.



Analyzing Big Data

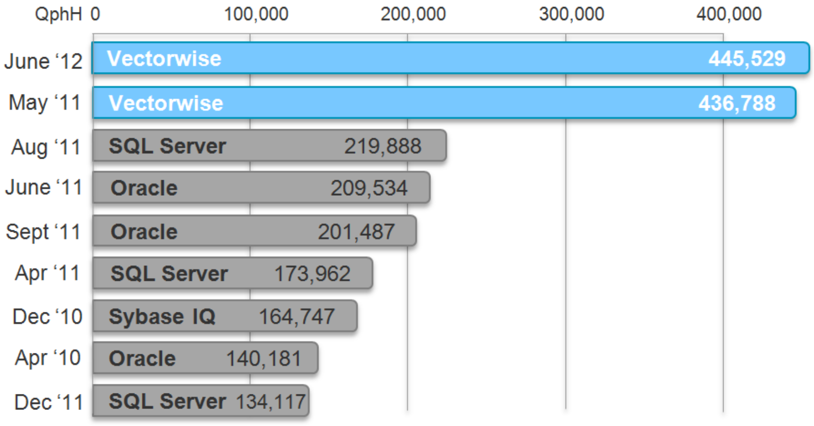
Hadoop is a massively parallel architecture that allows for the distributed processing of large data sets across clusters of computers. Hadoop is an architecture that is ideal for collecting, managing and retrieving Big Data, however it can be a painfully slow to query data. While Hadoop allows parallelization across all nodes in the cluster, the overhead introduced and the sheer volume and variety of data can result in long query times, especially when user interaction drives the analytics. As a result, Hadoop is not ideal for fast ad-hoc queries, let alone highly concurrent queries.

Vectorwise is the record-breaking analytic database for fast reporting and analytics on Big Data. Vectorwise is an ACID compliant, ANSI SQL-based relational database with a performance engine uniquely designed to exploit the performance features in today's x86 CPUs. As a result, Vectorwise can process data significantly faster than other relational databases, allowing organizations to analyze more data faster. In addition, Vectorwise enables you to run all workloads on a single server when other databases require a much larger machine, a cluster of servers, or both, to achieve similar results.

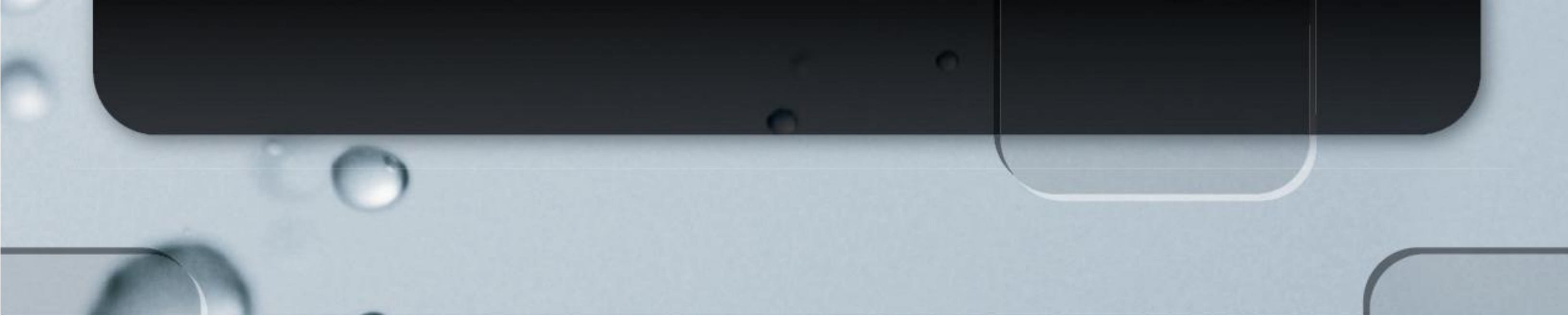
Vectorwise holds numerous TPC-H records for performance, price/performance and energy efficiency – exceeding previous records by the largest margins ever recorded (www.tpc.org/tpch).

Figure 1: Fastest TPC-H QphH@1TB Benchmark (non-clustered)

Source: www.tpc.org / March 15, 2013



Hardware Costs (Excluding Discounts)	Processors	Cores	RAM
$57,146	4	32	768 GB
$85,621	4	32	1 TB
$460,869	8	80	2 TB
$2,402,706	16	64	512 GB
$753,392	4	32	512 GB
$279,527	8	80	128 GB
$1,229,968	8	32	512 GB
$1,249,967	16	64	512 GB
$258,880	4	40	1 TB

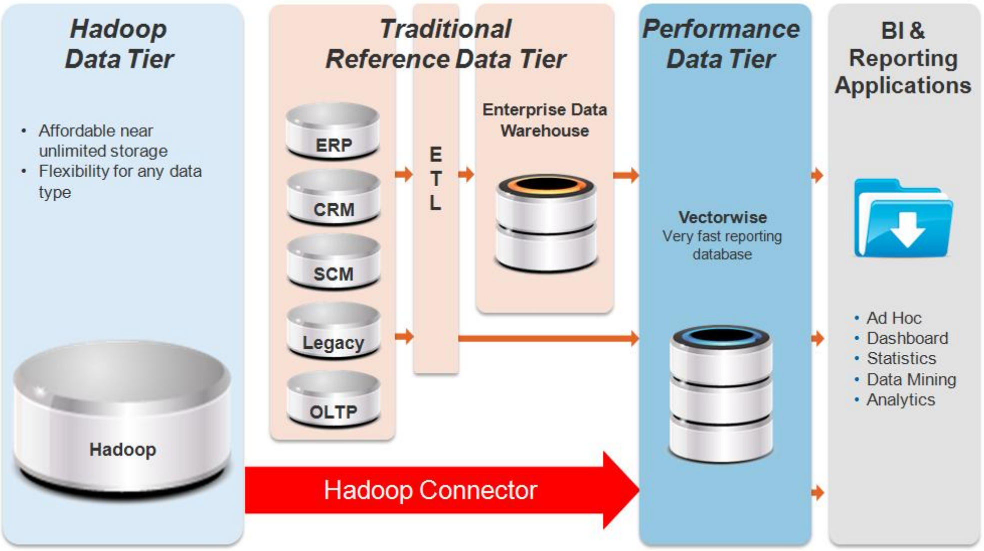


Vectorwise and Hadoop Environments

When combined, Vectorwise and Hadoop deliver a potent mix for solving Big Data analytics issues. Vectorwise is being used to boost the performance of analyzing Hadoop data in dozens of use cases from social media, to online gaming/dating, and data aggregators at companies such as NK, IsCool Entertainment and edo interactive.

Figure 2 below shows how Vectorwise can be used to speed up analytic queries in Hadoop and other systems. Vectorwise is able to provide fast answers for analysis between both relational systems and data stored in Hadoop.

Figure 2: Typical Vectorwise and Hadoop Environments


Hadoop Data Tier
Affordable near unlimited storage
Flexibility for any data type
Hadoop
Traditional Reference Data Tier
ERP
CRM
SCM
Legacy
OLTP
E T L
Enterprise Data Warehouse
Hadoop Connector
Performance Data Tier
Vectorwise
Very fast reporting database
BI & Reporting Applications
Ad Hoc
Dashboard
Statistics
Data Mining
Analytics

Vectorwise Hadoop Connector

The Vectorwise Hadoop Connector is the fastest and most cost-effective way to load massive volumes of data from Hadoop into Vectorwise.

The Vectorwise Hadoop Connector functions as a parallel bulk loader between Hadoop/HDFS and Vectorwise. The connector uses Hadoop's parallel processing capability to offload the data loading stage in Vectorwise by pre-building the compressed Vectorwise storage blocks of data. The net effect is having very little cycles spent on the Vectorwise database server, leaving cycles for fast SQL query processing on data.

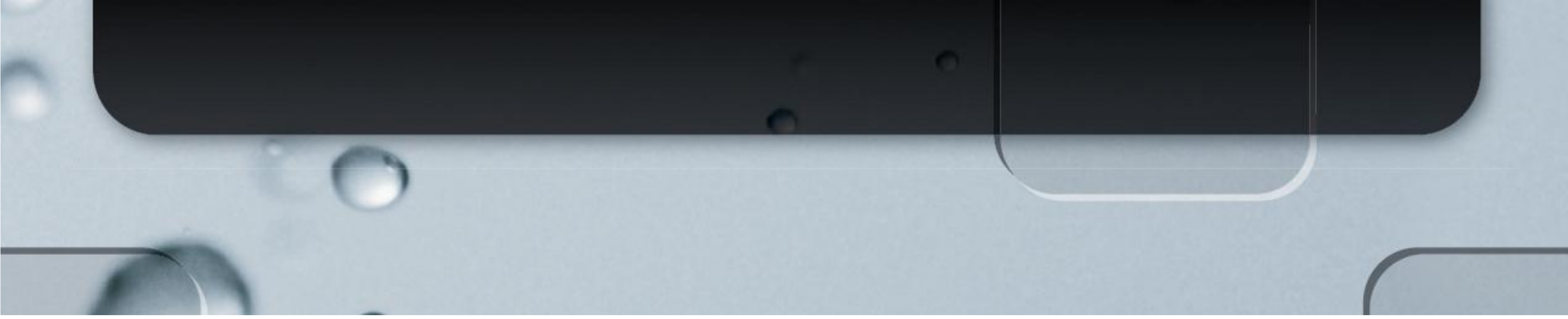
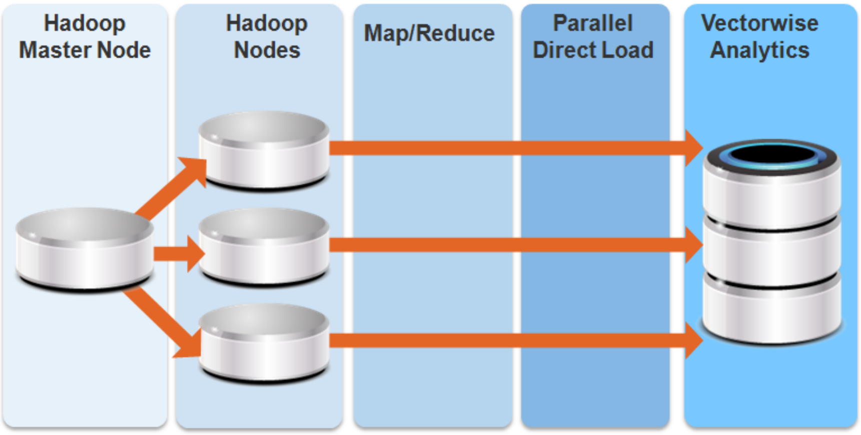


Figure 3: Vectorwise Hadoop Connector



Vectorwise Hadoop Connector is developed and supported by Actian, and is designed to work with stable Hadoop versions 1.0.3 or above.

Performance Tests

The performance tests measured load times for a variety of different data sizes ranging from 100GB to 1TB.

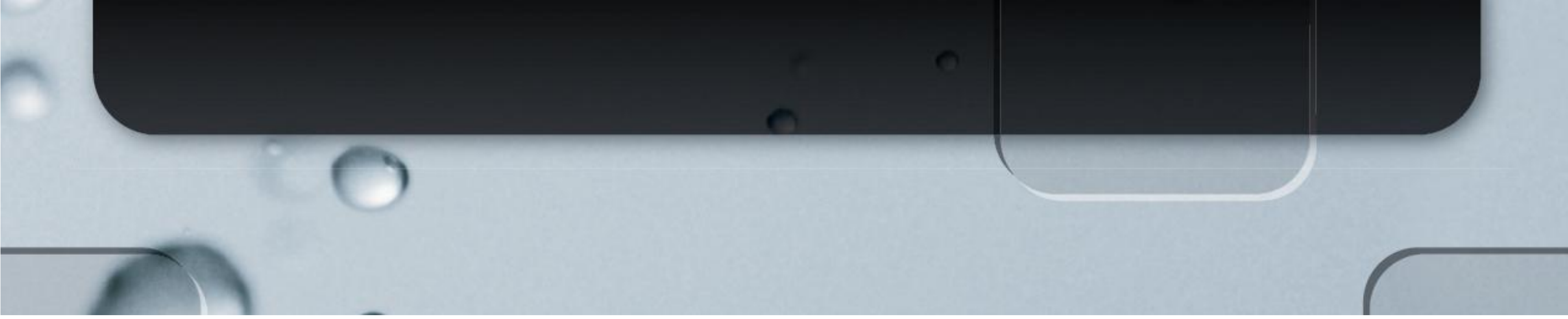
Configuration Overview

The data load rates depend on multiple factors including the Hadoop and Vectorwise hardware configuration, the data volume and types of data, and the table definitions.

A Dell R720 with 2x Intel E5-2650 CPUs (8 cores/CPU, 2GHz clock speed with turbo at 2.8GHz, 20 MB cache), 192 GB RAM, 12 x 300 GB 10k RPM drives was utilized for all tests.

In total, 5 identical servers were used in total, with 1 server running Vectorwise, and 4 servers for the small Hadoop cluster (1 master node and 3 slaves running Map/Reduce jobs). Each slave in the Hadoop cluster was configured to run a maximum of 35 Map tasks and 24 Reducer tasks.

The total hardware cost for the Vectorwise server was less than $15,000 USD.



Data Load Rates from Hadoop to Vectorwise

Data load sizes ranged from 115GB (18 million rows) to 1TB (158 million rows). The target table had 48 columns (2x integer columns, 1x narrow character column with 20 characters, 45x wide character-based columns with 170 characters). Data was generated completely at random.

Table 1 below shows that the load rate from when the reducer first sends data to the Vectorwise server varied between 2 TB/hour and 3 TB/hour. Peak load rate was achieved with 230GB loaded in just 4 minutes and 35 seconds at a rate of 3.03 TB/hour. 1TB loaded at a rate of 2.14TB/hour in 28 minutes 30 seconds.

Table 1: Vectorwise Hadoop Connector Performance Test

Data Load		Load Time	Data Load Throughput	Data Load Throughput
Rows	**Data size**	**Elapsed Time**	**GB / SEC**	**TB / HR**
18M	115.8 GB	2 min 21 sec	0.82	2.957
36M	231.6 GB	4 min 35 sec	0.84	3.032
54M	347.4 GB	8 min	0.72	2.606
72M	463.2 GB	11 min 12 sec	0.68	2.481
108M	694.8 GB	20 min 1 sec	0.57	2.083
158M	1016.96 GB	28 min 30 sec	0.59	2.141

All tests were performed with the same cluster configuration regardless of data set size. Figure 4 below shows that for larger data set sizes, 500GB and above, load performance stabilized around the 2TB/hour range. The results above were achieved with 3 Hadoop slaves.

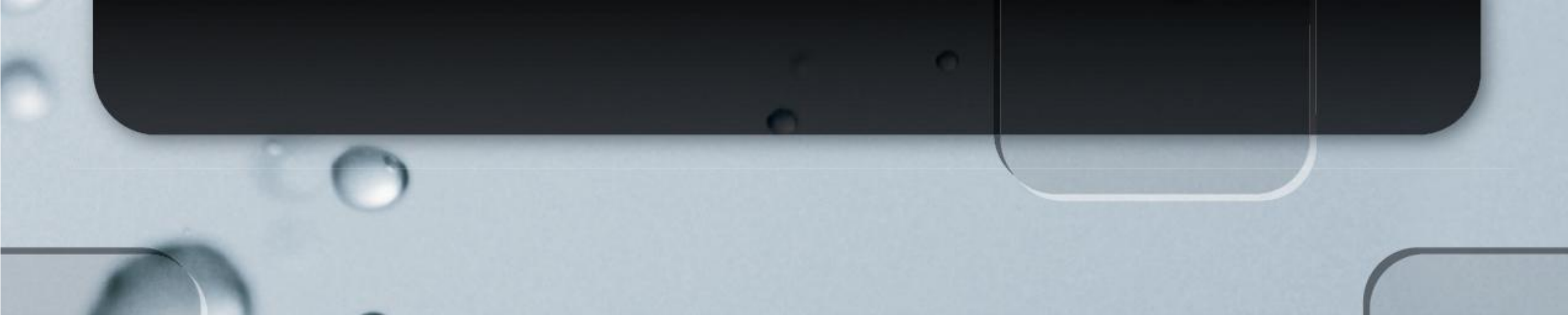
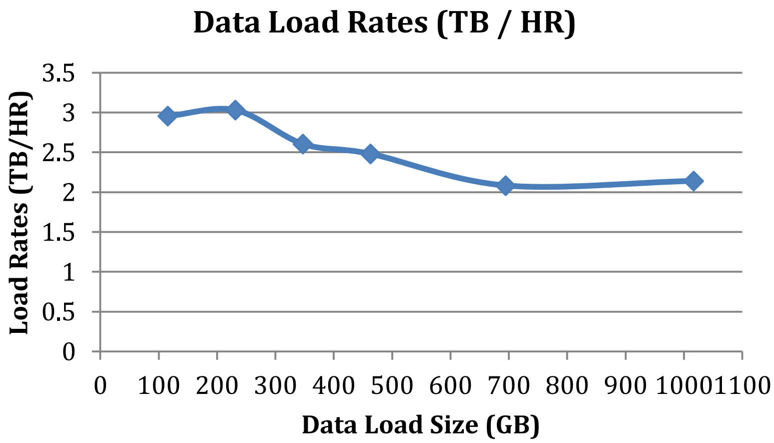


Figure 4: Comparing data load rate for different data sizes



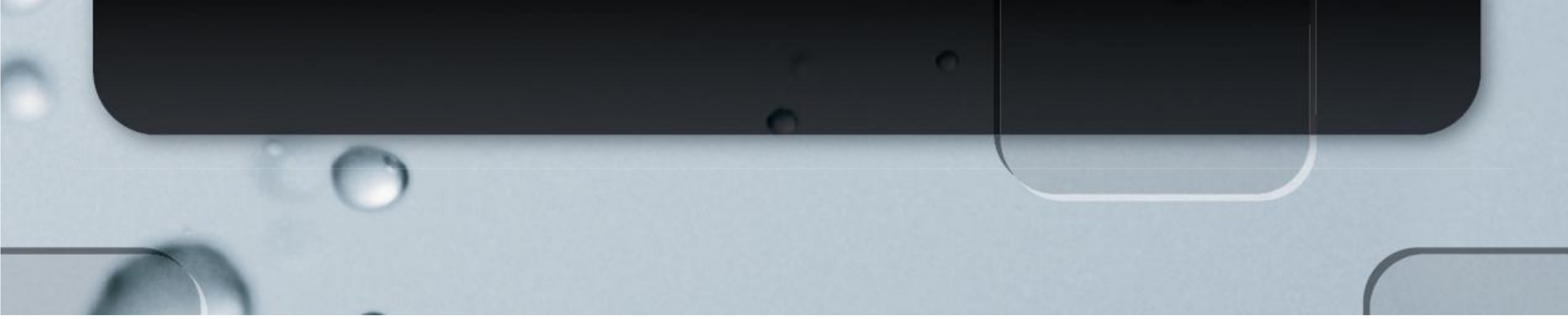
End-to-End Data Load Rates from Hadoop to Vectorwise

The tests below measured the end-to-end load time starting from when the Hadoop job starts on the small Hadoop Cluster (previous results measured from when the data started hitting the Vectorwise database). While the throughput rate is lower at between 0.59 TB/HR to 0.48TB/HR, this includes building the compressed Vectorwise data blocks in Hadoop before pushing the load into the Vectorwise database.

Table 2: Vectorwise Hadoop Connector End-to-End Performance Test

Data Load		End to End Load Time	Data Load Throughput	Data Load Throughput
Rows	**Data size**	**Elapsed Time**	**GB / SEC**	**TB / HR**
18M	115.8 GB	14 min 13 sec	0.136	0.489
158M	1016.96 GB	2 hrs, 12 min 56 sec	0.128	0.459

Adding more Hadoop slaves will further reduce the amount of time spent building data blocks in Hadoop, and therefore improve end-to-end throughput for the Vectorwise Hadoop Connector.



Conclusion

Hadoop and Vectorwise offer complementary solutions for solving Big Data problems. Hadoop offers an infinitely scalable data store able to collect, organize, store and retrieve large amounts of data. Vectorwise is a record-breaking database for Big Data analytics. The Vectorwise Hadoop Connector makes it fast and affordable to load data from Hadoop into Vectorwise.

The combined, multi-tiered architecture of Hadoop and Vectorwise allows users to analyze Big Data faster and cost-effectively on commodity hardware. The Vectorwise Hadoop loader allows seamless access to Hadoop data, enabling near real-time analysis for optimized decision making and taking action on Big Data.

For more information and to download Vectorwise and Hadoop Connector evaluation software, go to http://www.actian.com/vectorwise.

About Actian: Take Action on Big Data
Actian Corporation enables organizations to transform big data into business value with data management solutions to transact, analyze, and take automated action across their business operations. Actian helps 10,000 customers worldwide take action on their big data with Action Apps, Vectorwise the analytic database, Ingres an independent mission-critical OLTP database, and Versant object-oriented database. Actian is headquartered in California with offices in New York, London, Paris, Frankfurt, Hamburg, Amsterdam and Melbourne.

Actian Corporation
500 ARGUELLO STREET
SUITE 200
REDWOOD CITY
CALIFORNIA 94063
USA
PHONE: +1.650.587.5500

Actian Europe Limited
217 BATH ROAD
SLOUGH
BERSHIRE, SL 1 4AA
UNITED KINGDOM
PHONE: +44 (0) 17.5355.9500

Actian Germany GmbH
OHMSTRASSE 12
63225 LANGEN
GERMANY
PHONE: +49 (0) 6103.9881.0

WEIMARER STR. 1A
D-98693 ILMENAU
GERMANY
PHONE: +49 (0) 3677.6785.0

Actian France
IMMEUBLE GABRIEL VOISIN
79 RUE JEAN-JACQUES
ROUSSEAU
92150 SURESNES
FRANCE
PHONE: +33 (0) 1.80.03.11.50

Actian Australia
LEVEL 8, SUITE 1
616 ST. KILDA ROAD
MELBOURNE, VICTORIA, 3004
AUSTRALIA
PHONE: +61 3.8530.1700

ACTIAN.COM
FOR MORE INFORMATION, CONTACT ACTIAN@ACTIAN.COM