

# SOLVING BIG DATA CHALLENGES WITH RED HAT AND INTEL

## TECHNOLOGY OVERVIEW



Growing and evolving big data challenges require a flexible and scalable set of technologies.

The Intel Distribution for Apache Hadoop offers a supported enterprise-class Apache Hadoop platform, tuned and supported on Intel® Xeon® hardware platforms.

Red Hat® JBoss® Data Grid provides in-memory solutions that perform and



scale exceptionally on large memory Intel Xeon hardware platforms.

Together these technology solutions can be combined in innovative ways to solve a host of big data challenges.

## INTRODUCTION

Billions of connected user devices and new levels of sensor data are driving unprecedented data growth, and big data insights now truly represent a new frontier for innovation. Highly diverse organizations are trying to respond to dynamic challenges and opportunities by deploying big data solutions. Not only is the persona of data constantly changing in terms of data volume, velocity, and variety, but big data technology itself is rapidly shifting and evolving to enable new business models and deliver new timely insights. More than merely deriving benefit and value from their data, organizations now see big data technology as a unique opportunity to innovate and deliver critical real-time results, resiliency, and security. Hardware technology too is evolving to serve these new requirements, even as it enables new levels of performance and scalability.

Ultimately, organizations need a suite of big data tools to shape flexible and effective solutions in line with business and application requirements. Solutions must be able to accommodate both structured and unstructured data, or combinations of both. While batch processing of data fits some application models, real-time options are increasingly required to evaluate and process data as it arrives, with additional analytics providing added value. With memory prices dropping and individual server memory capacities extending into the terabytes, in-memory processing options are now often deployed in combination with traditional disk-based solutions. Most importantly, infrastructure to handle big data workloads must scale effectively to support the memory, processing, networking, and I/O requirements of these new distributed workloads and data.

With a strong track record of innovation and deep collaboration, Red Hat and Intel have supported mission-critical environments for years, helping organizations to optimize their cost models while maintaining the highest levels of performance and scalability. This extensive collaboration helps ensure that Intel processor technology works well with Red Hat operating system technology and other software. The collaboration and innovation extends to big data technology, and tools from both companies complement each other and help to form a more comprehensive solution for diverse application needs. Intel Xeon-based hardware platforms supply large memory capacities, high performance, and balanced I/O, serving as ideal infrastructure for hosting Red Hat® JBoss® Data Grid non-relational in-memory data storage. Likewise, the Intel Distribution for Apache Hadoop is engineered from the hardware throughout the entire stack, representing an ideal complement to the capabilities of Red Hat JBoss Data Grid technology.

## THE BIG PICTURE FOR BIG DATA

As most organizations quickly realize, exploiting big data opportunities requires an approach that goes beyond single technology solutions. Different data processing challenges require distinct ways of handling and interacting with data. Extending beyond distributed data and distributed processing, Figure 1 shows a big-picture perspective on the spectrum of technologies that may be required for big data analysis, with the capabilities of Red Hat JBoss Data Grid highlighted. The combination of Red Hat JBoss Data Grid with the Intel Distribution for Apache Hadoop (including Apache HBase and Apache Hive) covers most required big data capabilities. While not all of these capabilities will be required for every application, having access to a range of technologies is vital for effective problem



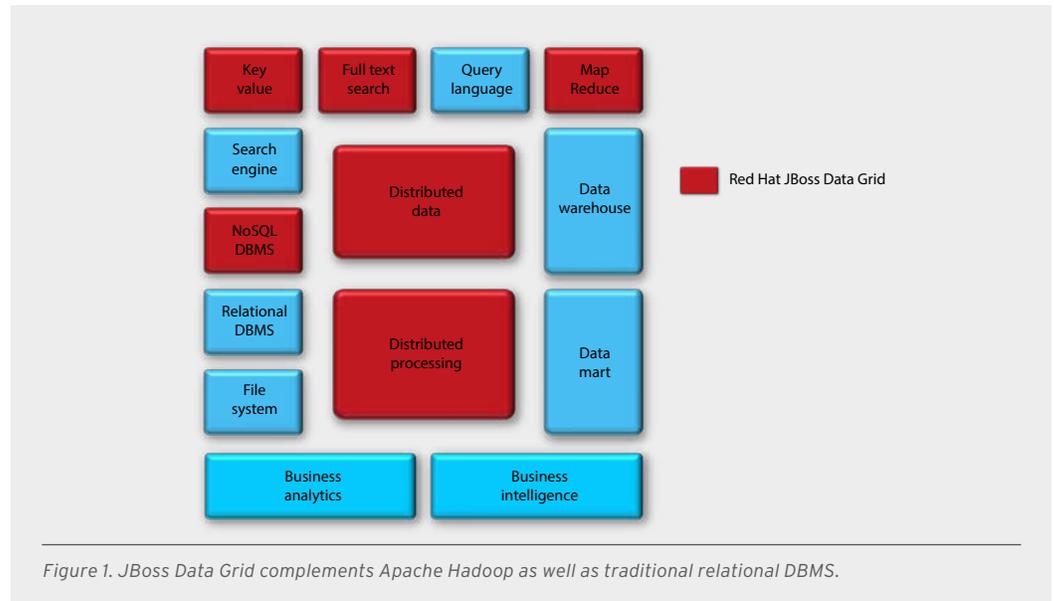
facebook.com/redhatinc  
@redhatnews  
linkedin.com/company/red-hat

By 2014, at least  
**40%**

of large organizations will have deployed one or more in-memory data grids

(Gartner Predicts 2012, Cloud and In-Memory Drive Innovation in Application Platforms)

solving in the modern big data universe. With these complementary technologies, both structured and unstructured data are easily supported. In-memory and disk-based approaches can also be combined to serve the broadest set of possible big data applications.



## INTEL DISTRIBUTION FOR APACHE HADOOP

Traditional distributions of Apache Hadoop have filled a valuable niche, but often fall short in terms of performance, quality, functionality, and enterprise-class features. In contrast, the Intel Distribution for Apache Hadoop (Intel Distribution) provides Apache Hadoop and other software components along with key enhancements and fixes from Intel. Proven in production at some of the most demanding enterprise deployments in the world, the Intel Distribution is supported by a worldwide engineering team with access to expertise throughout the entire software stack, as well as the underlying processor, storage, and networking components.

Key features of the Intel Distribution for Apache Hadoop include:

- Up to 30 times the Hadoop performance with optimizations for Intel Xeon processors, Intel® SSD storage, and Intel® 10 Gigabit Ethernet networking
- Data confidentiality without a performance penalty, with encryption and decryption in the Apache Hadoop Distributed File System (HDFS) enhanced by Intel® AES-NI and role-based access control with cell-level granularity in HBase
- Multi-site scalability and adaptive data replication in HBase and HDFS
- Up to a 3.5-fold performance improvement in Hive query performance
- Support for statistical analysis with R connector
- Graph analytics with Intel® Graph Builder
- Enterprise-grade support and services from Intel

## DEPLOYING RED HAT JBOSS DATA GRID

As a NoSQL key/value data store, Red Hat JBoss Data Grid provides the flexibility to store any type of data in a data element, at in-memory speeds. Easily added to applications and existing relational databases, Red Hat JBoss Data Grid also provides a number of distinct performance advantages, including:

- Easing the load on database servers
- Shortening response times and lowering latency in applications
- Providing failure resilience, and
- Offering data access via a variety of access protocols, including REST, memcached, Hot Rod, or a simple map-like API

Coupled with Apache Hadoop elements, Red Hat JBoss Data Grid can be used to drive a wide range of applications with diverse needs for scaling, latency, and redundancy. A few examples are described in the sections that follow.

- **Securities trading.** By writing simultaneously to a database and Red Hat JBoss Data Grid, low latency and real-time trade processing are easily achieved, backed up by writes to the relational database for complete transactions.
- **Video streaming.** Red Hat JBoss Data Grid is ideal for low-latency applications such as video streaming that need to track a related resource state.
- **Structured data.** Using Red Hat JBoss Data Grid, tiered storage principals can easily be applied to structured data—for example, storing a day’s worth of data in the data grid, a month’s worth of data in a database, and a year’s worth of data in Apache Hive.
- **Business analytics.** Business analytics applications can employ Red Hat JBoss Data Grid as a cache, reading data from Apache HDFS and storing results in the grid so that they can be accessed and interrogated to derive additional business value.
- **Stream processing.** Red Hat JBoss Data Grid can be an effective tool for stream processing, allowing analysis of trending events even as data is stored in parallel in Apache HBase.
- **Data logging.** Data logging can represent another appropriate use of Red Hat JBoss Data Grid, with log files cached via Hot Rod to the grid for real-time access, and then batched to Apache HDFS via Flume.

## RED HAT JBOSS DATA GRID ON SCALABLE INTEL® XEON® PLATFORMS

Intel Xeon platforms provide the preferred solution for in-memory analytic engines and scale-up databases such as Red Hat JBoss Data Grid. The platform provides columnar database performance improvements through the use of Intel AVX instructions. Essential for large in-memory databases, four-socket platforms provide support for up to two terabytes of memory. These systems also provide very high reliability and full eight-socket scalability.

To evaluate the performance of Red Hat JBoss Data Grid on Intel Xeon platforms, a configuration was built using three physical servers, each equipped with four Intel Xeon E7-4860 processors, a terabit of RAM, and 10 Gigabit Ethernet networking. The resulting 30-node in-memory Red Hat JBoss Data Grid was able to provide over 140,000 transactions per second and over 514,000

Contact your Red Hat Sales Representative to learn more about running Red Hat JBoss Data Grid on Intel Xeon platforms.

reads per second (Table 1). It is important to note that these results are highly conservative, and were derived with full data object replication, fully enabled transactions, and synchronous communications<sup>1</sup>.

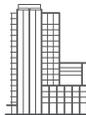
Table 1. Intel Xeon platforms provide excellent scalability for Red Hat JBoss Data Grid.

Physical servers	3
Intel Xeon E7-4860 processors	12
RAM	3 TB
Red Hat JBoss Data Grid nodes	30
Data	1+ TB (1,035 GB)
Transactions per second	140,172
Reads per second	514,168
Writes per second	55,181

## CONCLUSION

Since the advent of Apache Hadoop, big data challenges and opportunities have continued to grow. Organizations today need access to a complete arsenal of big data technologies in order to produce innovative and effective big data solutions that address changing needs and new opportunities. The strong relationship between Red Hat and Intel around big data means that tools like Red Hat JBoss Data Grid combined with the Intel Distribution for Apache Hadoop provide effective enterprise-grade solutions. Moreover, Intel Xeon hardware provides an ideal platform for Red Hat JBoss Data Grid, helping to ensure performance and scalability for a wealth of diverse big data applications.

1. Preliminary scalability testing conducted by Red Hat and Intel. Actual performance may vary depending on a variety of factors and conditions.



## ABOUT RED HAT

Red Hat is the world's leading provider of open source solutions, using a community-powered approach to provide reliable and high-performing cloud, virtualization, storage, Linux, and middleware technologies. Red Hat also offers award-winning support, training, and consulting services. Red Hat is an S&P company with more than 70 offices spanning the globe, empowering its customers' businesses.



facebook.com/redhatinc  
@redhatnews  
linkedin.com/company/red-hat

**NORTH AMERICA**  
1-888-REDHAT1

**EUROPE, MIDDLE EAST  
AND AFRICA**  
00800 7334 2835  
europe@redhat.com

**ASIA PACIFIC**  
+65 6490 4200  
apac@redhat.com

**LATIN AMERICA**  
+54 11 4329 7300  
info-latam@redhat.com