EXECUTIVE SUMMARY

Data scientists use big data to analyze data that was too large or too difficult to organize conventionally. Using distributed scaling, big data analysis can deliver timely results and value to the corporation or agency. Big data is a complex application environment requiring advanced system management and superior data throughput. Because big data solutions quickly become business-critical, enterprises and agencies require an operating platform that is reliable, secure, high-performance, open, and flexible, with engineering support aimed at delivering the latest industry innovations.

To avoid bottlenecks in the big data process, IT architects and data scientists need to choose an operating environment that can grow from pilot to production. Understanding the ecosystem, application profiles, and underlying technologies is a prerequisite for anyone undertaking a big data project.

With scalable production tools and a rich development environment, Red Hat® Enterprise Linux® delivers capabilities for both IT operations and data scientists. Standardizing on Red Hat Enterprise Linux reduces complexity, improves performance, and accelerates value.
THE BIG DATA TRANSFORMATION

Every enterprise has a wealth of digital information from their regular interactions with customers, finance, manufacturing, and logistics. Typically, these data silos operate independently. Combining them into a single environment is technically difficult because each data silo is different and the overall volume is unwieldy.

Big data technology has made storing and querying large, disparate data sets feasible. Big data solutions embrace the variety of data available to create a rich context for strategic advantage and improved tactical response. Big data spreads work across multiple systems to eliminate common database bottlenecks. Scaling out requires dynamic data models and a flexible application framework.

A Linux-based platform is the clear choice to support the scalable open source Java™ framework on which most big data solutions are built. Red Hat Enterprise Linux is well-known for superior performance, scalability, and reliability for traditional databases, it is also the logical choice for big data projects that need to run the latest big data technologies.

OPEN SOURCE CHANGED THE GAME

Hadoop, the most common software associated with big data, was originally developed by Yahoo for their large Linux datacenter. Hadoop pushes the computation to the distributed data. Adding servers to the cluster increases both capacity and computational power with a technique called MapReduce. Created to solve the specific problem of data analysis across many servers, Hadoop has changed the industry.

Hadoop has always been open source software. Primarily written in Java, the components of Hadoop framework have benefited from the support of a broad community. Contributions come in from around the globe, written and tested by individual developers, users, academics, and companies of all sizes.
**OPEN SOURCE CHALLENGES**

Open source easily supports multiple, related development projects to drive innovation. However, working from a common codebase does not ensure compatibility. It falls to the IT professional to manage applications, libraries, and systems.

Red Hat is the leader in open source engineering, with decades of experience integrating open source, system hardware and commercial software into a high-performance, secure platform.

---

### THE HADOOP STACK

There is an entire lexicon of open source data analytics tools used in big data clusters. The core technologies – Hadoop Distributed File System (HDFS), Yarn, and MapReduce – manage data and job distribution across the servers. Data queries are built from the layered utilities included with Hadoop, tailored for different data types. Hadoop queries are written as multistep scripts using several utilities in combination.

The Hadoop stack includes tools for job scheduling, monitoring, and cluster management. However, IT architecture expertise is needed to analyze infrastructure performance and set resource alerts that can be used for planning upgrades and cluster growth.

#### DATA TRANSFER

- Flume
- Sqoop
- Database

#### DATA ANALYSIS

- Pig
- Hive
- Mahout
- HBase
- Distributed compute (YARN)
- Distributed file system (HDFS)

#### CLUSTER AND JOB MANAGEMENT

- Ambari
- Falcon
- Oozie
- Zookeeper

---

Figure 1. The Hadoop Platform

Hadoop and related projects are freely available from hadoop.apache.org or as a stable release with commercial support from several companies including Cloudera, Hortonworks, and Intel.

Production platforms should be balanced for Java performance and multi-application efficiency. The data access patterns on the cluster will be dependent upon the query and the tools used. Data can be shared in memory, on local storage or between nodes. The system should provide fast access to large local file systems and high-speed networking across the cluster.
ALL THE DATA

By definition, big data projects involve multiple data sources. Hadoop may be ideal for unstructured data, but most business data is, in fact, structured. Traditional databases are faster for most transactions at the core of daily business operations.

Data transfer time can be a large component of overall processing time for big data queries. Pre-processing the data will reduce transfer volume, but is not as flexible as processing the data on the big data cluster. Traditional database vendors, such as Oracle and IBM, include Hadoop connectors to move data between their relational databases and Hadoop’s distributed environment.

Many organizations obfuscate or remove sensitive information, such as credit card numbers or patient names, during data movement. In addition, complex event processing (CEP) functions can be used to extract insight in real time from the data stream. Other applications pre-process the data using extract, transform, and load (ETL) semantics.

Architects and operations need to effectively plan and actively manage these data transfers. Network, compute, and storage should have established service levels for data transfer, as well as regular reporting and trend analysis.

Deploying Red Hat Enterprise Linux in these environments provides multiple benefits and allows effective resource allocation. Red Hat Enterprise Linux can allocate compute, I/O, and memory resources per virtual environment, application, or thread. These can be readily incorporated into daily operations for maximum operational flexibility.

![Figure 2. Big data is a part of the enterprise data portfolio](image_url)

Big data projects do not stand alone; they bring together data and analysis from across the organization. Technical architects need a broad view to anticipate the many places a big data project will impact the business and the IT infrastructure.
Supporting DevOps

Rapid innovation by the community and commercial vendors makes big data kinetic. Data scientists test new technologies and validate different techniques daily. The platform must support a multiplicity of development, test and production environments seamlessly.

Using the expertise of a commercially supported Linux platform will simplify migration from pilot projects to a demanding production environment.

Standardizing on Red Hat Enterprise Linux ensures consistency and security across containers, virtual machines, physical systems, and cloud providers.

INTELLIGENCE FOR A DYNAMIC ENVIRONMENT

A big data project is dynamic in development and data growth. To maximize development, application and deployment flexibility, while also delivering performance and reliability, IT architects should standardize on a single operating system for big data and related initiatives.

DYNAMIC DEVELOPMENT

Platform consistency is essential to smooth migration from development to production. There is a natural tension between the data scientist and the IT architect. Data scientists want to use the latest applications, though they may affect system stability and security. IT architects favor stable enterprise Linux platforms because they save time, remove complexity, and improve reliability.

Make sure that big data projects have enterprise capabilities and the latest technology with Red Hat Developer Toolset and Red Hat Software Collections. Red Hat packages the latest stable development tools and run-times for dynamic development environments. These packages are thoroughly tested and supported by Red Hat and allow IT departments to track the fast pace of big data tool development without compromising security or platform stability.

DYNAMIC USER SPACE

Data scientists need tools to efficiently run multiple applications on the clusters. Linux containers allow environment isolation and resource control, both invaluable in big data.

Unlike hypervisor-based virtual machines, Linux containers are built on top of a single instance of an operating system. They run on the same kernel, but they’re shielded from other containers. Users with sufficient privileges can run multiple instances of the same application or different applications with different libraries.

Linux containers can be quickly deployed for testing or incorporated into production scripts. In testing, a Linux container provides a dedicated environment for the data scientist to develop without the complexity associated with system installation and configuration. In production, Linux containers can host multiple Java virtual machines (JVMs) on a common piece of hardware, increasing system parallelism.

Red Hat Enterprise Linux can assign resource levels to different containers on the same machine. Memory, compute, and I/O resources can be assigned when spawning a container, or adjusted dynamically. In this way, query resources can be optimized on each system. Balancing and isolating multiple applications reduces contention and improves system throughput.

DYNAMIC PROVISIONING

Big data workloads follow the cadence of business, with different requirements by time of day, month and, year.

Dynamic provisioning and management is made easier with Red Hat Enterprise Linux. The Kernel-based virtual machine (KVM) provides a consistent environment across cloud, virtual machines, and physical deployments. As big data projects grow, standard system images and scalable update processes are critical.

Red Hat Enterprise Linux is a foundation for an agile big data project. It provides the technology and control to support code development, multiple simultaneous queries, and time-of-day flexible provisioning.
OpenJDK
OpenJDK is the leading open source JVM implementation. Hadoop projects are regularly developed and tested on it. Led by Red Hat and with many contributors, OpenJDK is a mature, high-performance environment for big data.

Red Hat Enterprise Linux ships with the latest version of OpenJDK. The operating system and OpenJDK are both optimized to deliver superior I/O performance.

BIG DATA PERFORMANCE CONSIDERATIONS
Hadoop’s distributed architecture provides system-level fault tolerance and scalable performance. However, it is layered on a standard server environment. Compute throughput, file-system performance, and network reliability are determined by the choice of the underlying server platform.

A modern enterprise server has 16 or more processor cores, several hundred gigabytes of memory, and multiple high-speed network interfaces. Proper allocation of processors, memory, and I/O is a function of the operating system and influences performance.

COMPLEX COMPUTE
Big data solutions typically include a collection of Java tools running as a number of interrelated jobs. Each job runs in its own JVM, and it is common practice to run more JVMs than processor cores. Oversubscription reduces the I/O wait impact and improves system throughput.

The operating system manages processes and memory affinity. The NUMA (non-uniformed memory access) architecture of modern systems imposes a penalty to access, move, and copy data between CPUs within the system. Allocating memory that is directly attached to the execution core lowers latency and improves performance.

Kernel algorithms use the system mapping of processor core and memory location to tune memory placement. Red Hat works closely with AMD, Intel, and server original equipment manufacturers (OEMs) to optimize Red Hat Enterprise Linux and certify all their servers. Red Hat Enterprise Linux supports the latest architectures and instruction sets and leads the Linux community in performance benchmarking. Compute performance and multi-core scalability are assured when using Red Hat Enterprise Linux.

BIG DATA I/O
Hadoop works on distributed, unstructured data and requires high throughput across different data types. The operating system manages physical storage, local system performance, and data integrity.

Random data distribution across the cluster spreads out the Hadoop workload, but is a liability in local storage performance. Random reads and writes require jumping across disk sectors and often defeat common disk performance enhancements such as read-ahead and caching.

The use of direct-attached storage with multiple disks is often chosen for big data challenges to lower costs and reduce storage contention. Red Hat Enterprise Linux uses either ext4 or XFS® file systems for local storage. Both are high-performance, journaled file systems with advanced enterprise features. Journaling is critical for large file systems across multiple disks, where it greatly improves rebuild and recovery time. Both file systems are engineered, tested, and tuned by Red Hat for excellent random I/O performance at scale.

Organizations that choose networked storage for big data will need support for the latest drivers and protocols. Red Hat is a leader in open source protocols development and industry storage support. Working alongside proprietary storage vendors, Red Hat Engineering has helped build protocol standards, such as Fibre Channels over Ethernet (FCoE) and Internet Small Computer System Interface (iSCSI). Red Hat Enterprise Linux was the first to support parallel-NFS, a new standard in clustered storage.
ENTERPRISE BIG DATA

Big data projects scale by adding systems to the cluster and naturally spawn multiple development and test environments. System provisioning, updates, and security can be complex. Using common tools throughout simplifies operations and encourages consistency.

Red Hat has the proactive engineering, scalable systems, and global footprint to support business-critical deployments of all sizes. System management scales using online and onsite tools. Red Hat’s knowledge base, system updates, and security notifications keep enterprises and agencies secure and stable at any size.

In addition to volume storage, some workloads, such as complex event processing, require low-latency random I/O performance only available from solid-state drives (SSD) based on flash memory. Peripheral Component Interconnect Express (PCIe)-based or external flash systems can be used to extended system memory or act as a cache in a tiered model. Red Hat Enterprise Linux is often used by flash vendors for their in-house development and testing to meet the needs of most demanding customers.

Red Hat publishes detailed tuning recommendations and tips based on real workloads and their engineering expertise. Red Hat Enterprise Linux further simplifies the tuning process with system profiles optimized for different compute and storage usage scenarios.

HIGH-SPEED NETWORKING

Due to the high volume of data ingested from other data sources and the data traffic within a big data cluster, big data clusters require a high-bandwidth, low-latency network architecture.

The time to move data to or from the big data cluster is primarily a function of bandwidth, but not just at the edge of the cluster. During data ingestion, the Hadoop file system creates pointers to the data and initiates copy protection between the nodes. Therefore, total network traffic is greater than just the copy from the original source. Isolating the traffic within the cluster is a good practice to simplify troubleshooting and reduce network contention.

Hadoop MapReduce pushes compute to the data, but does not protect the system from returning large results. Communication and comparison phases of the query can be latency or bandwidth limited. Only through analysis can it be determined if rewriting the query or upgrading the network is required.

10GbE networking should be standard in organizations with big data, but some clusters require multiple links either bonded together for bandwidth, or on separate vLANs to avoid contention. Advanced network protocols, such as RDMA over Converged Ethernet (RoCE), or virtual NICs to dedicate network devices to specific big data jobs, should also be considered.

Organizations handling big data have benefited from significant improvements in commodity servers and high-speed networking. More powerful commodity servers with high-throughput components and large file systems require an operating system developed alongside and tuned for those platforms. Red Hat Enterprise Linux maximizes reliability, performance, and hardware choice for big data clusters.
CONCLUSION

Big data projects are an opportunity for architects and data scientists to work together from the early planning stages. Choosing an open source platform that will support the evolution from a pet project to a business-critical big data solution is an important first step that should always be made cross-organizationally.

Red Hat Enterprise Linux is the leading platform for commercial and open source hardware and software development. It is the foundation to ensure security and integrity of enterprise data. Red Hat Enterprise Linux and its rich ecosystem provide the tools to build dynamic platforms capable of high productivity and scale. With support for the latest development environments, flexible job management, and consistent provisioning, standardizing on Red Hat Enterprise Linux allows big data innovation to thrive.

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.