Redbooks
ibm.com/redbooks

IBM Spectrum Scale

Big Data and Analytics Solution

Wei G. Gong

Sandeep R. Patil

Analytics

Storage

IBM

Redguide

IBM Spectrum Scale with Big Data and Analytics Solutions

IBM® Spectrum Scale is flexible and scalable software-defined file storage for analytics workloads. Enterprises around the globe deployed IBM Spectrum™ Scale to form large data lakes and content repositories to perform High Performance Computing (HPC) and analytics workloads. It is known to scale performance and capacity without bottlenecks.

Hortonworks Data Platform (HDP) is a leader in Hadoop and Spark distributions. HDP addresses the needs of data-at-rest, powers real-time customer applications, and delivers robust analytics that accelerate decision making and innovation.

IBM Spectrum Scale™ solves the challenge of explosive growth of unstructured data against a flat IT budget. IBM Spectrum Scale provides unified file and object software-defined storage for high-performance, large-scale workloads that can be deployed on-premises or in the cloud.

IBM Spectrum Scale includes NFS, SMB, and Object services and meets the performance that is required by many industry workloads, such as technical computing, big data, analytics, and content management. IBM Spectrum Scale provides world-class, web-based storage management with extreme scalability, flash accelerated performance, and automatic policy-based storage tiering from flash through disk to the cloud, which reduces storage costs up to 90% and improves security and management efficiency in cloud, big data, and analytics environments.

IBM Elastic Storage™ Server is an optimized disk storage solution that is bundled with IBM hardware and innovative IBM Spectrum Scale RAID technology (based on erasure coding). It performs fast background disk rebuilds in minutes without affecting application performance. Traditional Hadoop analytics systems feature a dedicated cluster to run Hadoop services, as shown in Figure 1 on page 2.

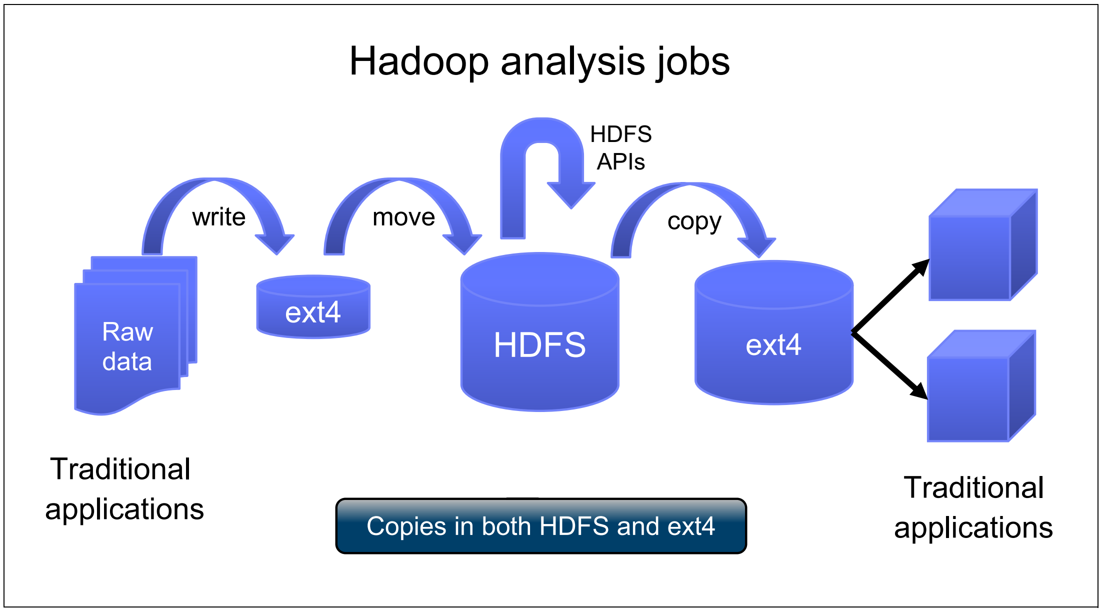

Figure 1 Data copy in traditional Hadoop analytics system

Data scientists waste days just copying data into Hadoop Distributed File System (HDFS) for analytics and copying data out of HDFS for traditional applications.

In this kind of analytics system, users and system administrators face the following key challenges:

- They need to build an analytics system from scratch for compute resource and storage.
- The inability to disaggregate storage resources from compute resources. To add storage capacity in the form of data nodes, an administrator must add processing and networking, even if they are not needed. This coupling of compute and storage limits an administrator's ability to apply automated storage tiering to use hybrid solid-state-drives (SSDs) or rotating disk architectures.
- Hadoop HDFS does not support the industry standard POSIX access interface. Therefore, to manage data, users must import data from systems, such as database and file store systems, to a Hadoop analytics cluster. Then, the analyzed result is exported back to another system (as shown in Figure 1). Data I/O processes can take longer than the actual query process.
- Many Hadoop systems lack enterprise data management and protection capability, such as data lifecycle management.
- A highly skilled system administrator is required to maintain HDFS.
- Data stability is highly affected by any server down in the cluster.

IBM Spectrum Scale is POSIX compatible, which supports various applications and workloads. By using IBM Spectrum Scale HDFS Transparency Hadoop connector, you can analyze file and object data in-place with no data transfer or data movement. Traditional systems and the analytics systems are using and sharing data that is hosted on IBM Spectrum Scale file system, as shown in Figure 2 on page 3.

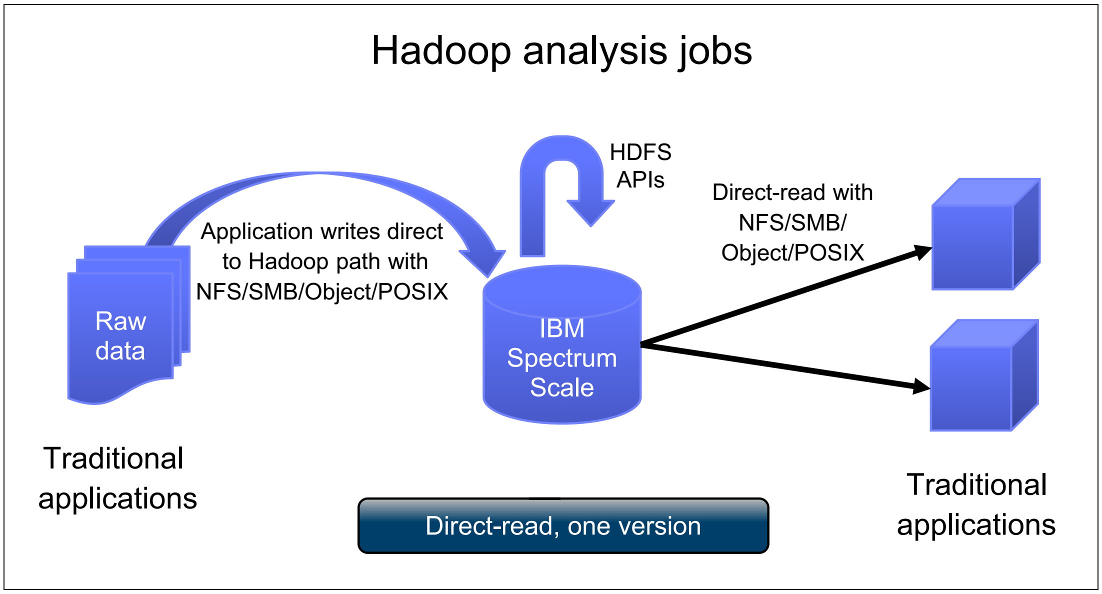

Figure 2   No data copy required in IBM Spectrum Scale analytics solution

Hadoop services can use a storage system to save IT cost because no special purpose storage is required to perform the analytics. IBM Spectrum Scale features a rich set of enterprise-level data management and protection features, such as snapshots, information lifecycle management (ILM), compression, and encryption, which provide more value than traditional analytic systems.

IBM Spectrum Scale for Spark

Spark is a fast and general engine for large-scale data processing. It runs programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

IBM Spectrum Scale supports for in-place analytics by using Spark, which can be configured with POSIX interface through IBM Spectrum Scale native client to mount the file system directly. It also supports the use of the HDFS interface through IBM Spectrum Scale HDFS Transparency.

IBM Spectrum Scale also supports various Spark distributions, such as Spark service in Hortonworks Data Platform and IBM Spectrum Conductor™ with Spark.

IBM Spectrum Scale features and benefits for big data and analytics

The following benefits can be realized by using IBM Spectrum Scale with HDP:

- Extreme scalability with parallel file system architecture

  IBM Spectrum Scale is a parallel architecture. With a parallel architecture, no single metadata node can become a bottleneck. Every node in the cluster can serve data and metadata, which allows a single IBM Spectrum Scale file system to store billions of files. This ability enables clients to grow their HDP environments seamlessly as the amount of data grows.

- Global namespace can span multiple Hadoop clusters and geographical areas

  By using IBM Spectrum Scale global namespace, clients can create active, remote data copies and enable real-time, global collaboration. This feature allows global organizations to form data lakes across the globe, and host their distributed data under one namespace. IBM Spectrum Scale also enables multiple Hadoop clusters to access a single file system while still providing all the required data isolation semantics.

  The IBM Spectrum Scale Transparent Cloud Tiering feature can archive data into S3/SWIFT-compatible cloud Object Storage system, such as IBM Cloud Object Storage or Amazon S3 by using the powerful IBM Spectrum Scale ILM policies.

- Data center footprint is reduced by using the industry's best in-place analytics

  IBM Spectrum Scale features the most comprehensive support for data access protocols. It supports data access that uses NFS, SMB, Object, POSIX, and the HDFS API. This feature eliminates the need to maintain separate copies of the same data for traditional applications and for purposes of analytics.

- True software-defined storage is deployed as software or a pre-integrated system

  You can deploy IBM Spectrum Scale as software directly on commodity storage-rich servers that are running the HDP stack. It also can be deployed as part of a pre-integrated system that uses the IBM Elastic Storage Server. Clients can use software-only options to start small and still use enterprise storage benefits. By using the IBM Elastic Storage Server, clients can control cluster sprawl and grow storage independently of the compute infrastructure.

- IBM hardware advantage

  IBM Elastic Storage Server uses erasure coding, which eliminates the need for the three-way replication for data protection that is required with other solutions. IBM Elastic Storage Server requires only 30% extra capacity to offer similar data protection benefits.

  Along with IBM Elastic Storage Server, IBM Power Systems™ offers the most optimized hardware stack for running analytics workloads. Clients can enjoy up to 3x reduction of storage and compute infrastructure by moving to Power Systems and IBM Elastic Storage Server compared to commodity scale-out x86 systems. IBM Elastic Storage Server also is available in all flash models, which allows for accelerated analytics that are required for certain data sets.

To support security and regulatory compliance requirements of organizations, IBM Spectrum Scale offers Federal Information Processing Standard (FIPS) compliant data encryption for secure data at rest, policy-based tiering and ILM, cold data compression, disaster recovery, snapshots, backup, and secure erase. The HDP Atlas and Ranger components provide more data governance capabilities and the ability to define and manage security policies.

IBM Spectrum Scale and HDFS comparison

In addition to comparable or better performance, IBM Spectrum Scale provides more enterprise-level storage services and data management capabilities, as listed in Table 1.

Table 1 Comparison of IBM Spectrum Scale (with HDFS Transparency) with HDFS

Capability		**IBM Spectrum Scale (with HDFS Transparency)**	**HDFS**
In-place analytics for file and object		Yes. All in place with support for POSIX, NFS, SMB, HDFS, and Object with concurrent access. Enables centralized, enterprise-wide data lakes.	Limited support with NFS gateway. No support for SMB, Object, or POSIX.
Performance		Comparable or better performance than HDFS in equivalent hardware configurations.	Same as IBM Spectrum Scale HDFS transparency.
Scalability (maximum number of nodes, files, and data)		IBM Spectrum Scale includes parallel file system architecture that differs from scale-out architecture of HDFS. No single metadata server is in the architecture as a bottleneck. Metadata serving function is distributed across the cluster. Test limit for number of files per file system is 9 billion. IBM Spectrum Scale production deployments are available beyond this test limit.	HDFS can scale up to 350 million files with a single name node because of scale-out architecture limitation. Supports only single or a pair of high availability NameNode, which becomes a bottleneck. Users must use federation functions to overcome this limitation.
If centralized storage is supported, what are the advantages?		Yes. Supports storage area network (SAN)-based shared storage and IBM Elastic Storage Server.	Not supported.
Supports storage-rich server		Yes	Yes
Supports tiering to tape and cloud Object Storage		Yes	No
Data reliability by using replication and erasure coding		Erasure codes from IBM Spectrum Scale RAID in IBM Elastic Storage Server, or data replication from IBM Spectrum Scale.	Support data replication for workload and erasure code for cold data.
Supports enterprise data backup		Yes, with IBM Spectrum Protect™ and Veritas NetBackup.	Does not support IBM Spectrum Protect or Veritas NetBackup.
Supports disaster recovery		Yes, Sync or ASync mode.	Only available for Hbase or Hive.
Supports Remote Direct Memory Access (RDMA)		Yes, when hardware is available.	Not supported.
Improve I/O performance through native client in compute node (Short Circuit Read/Write)		Yes, supports. Can use IBM Spectrum Scale Native Client and high-performance network, such as RDMA over InfiniBand to improve I/O bandwidth and latency and reduce CPU resource.	No native client on compute node.
Security	Secure data at rest	Yes, supports IBM ISKLM and Vormetric key manager and is FIPS-complaint,	Yes
	Secure data in motion	Yes	Yes
	Immutability	Yes	No
	Authentication	Yes	Yes
	Authorization	Yes	Yes
	Auditing	Yes	Yes
Ambari integration		Yes	Yes

When to consider IBM Spectrum Scale for big data and analytics solution

IBM Spectrum Scale for big data and analytics solution often is used for the following reasons:

- When the analytics system must collaborate with other traditional application systems that must use POSIX access interface.
- When HDFS-based Hadoop analytics systems encounter performance, stability, or scalability issues on the HDFS metadata node. IBM Spectrum Scale solves this problem because a centralized metadata node bottleneck does not occur and it is a proven enterprise level software-defined storage system.
- When your analytics system requires a global namespace that can span geographical areas.
- When you want to use industry-leading erasure code instead of data replication to protect data, which reduces the total cost of ownership.
- If you started your analytics system that uses commodity storage-rich servers, and now want to seamlessly and easily expand to use enterprise-level storage software and hardware.
- When you need a true software-defined storage solution to overcome HDFS scalability and availability limitations.
- When you need enterprise-class storage features to protect, archive, or back up your data.

Summary

IBM Spectrum Scale is a preferred platform for running big data and analytics workloads. IBM Spectrum Scale in-place analytics for file and object data solves traditional analytics solution challenges.

HDFS Remote Procedure Call (RPC)-based IBM Spectrum Scale HDFS transparency Hadoop connector provides enhanced high availability capability, performance, and security for big data and analytics workloads. The IBM Spectrum Scale big data and analytics solution is deployed in Storage Rich Server architecture, SAN shared storage, and an integrated system Elastic Storage Server.

POSIX compatibility with various enterprise class features provides flexible data management and protection for big data and analytics workloads.

For more information

For more information, see the following resources:

- IBM Knowledge Center for IBM Spectrum Scale:

  https://www.ibm.com/support/knowledgecenter/STXKQY/ibmspectrumscale_welcome.html

- Hortonworks Data Platform with IBM Spectrum Scale: Reference Guide for Building an Integrated Solution:

  http://www.redbooks.ibm.com/abstracts/redp5448.html

- Deploy a Big Data Solution on IBM Spectrum Scale:

  https://ibm.biz/BdiVKs

- IBM Spectrum Scale HDFS Transparency main page:

  https://ibm.biz/BdrY82

- IBM Spectrum Scale FPO deployment and maintain guide:

  https://ibm.biz/BdHKQF

- Hadoop from Sandbox to Production:

  https://www.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=DCL12414USEN

- Hortonworks documentation:

  https://hortonworks.com/partner/ibm

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new IBM Redbooks® publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

Redbooks (logo) ®
IBM®
IBM Elastic Storage™
IBM Spectrum™
IBM Spectrum Conductor™
IBM Spectrum Protect™
IBM Spectrum Scale™
Power Systems™
Redbooks®

The following terms are trademarks of other companies:

Other company, product, or service names may be trademarks or service marks of others.

REDP-5397-01

ISBN 0738456632

Printed in U.S.A.