Performance and Capacity Implications for Big Data

Overview and benefits of big data

Critical factors in managing big data successfully

Emerging technologies that address big data challenges

Dave Jewell
Ricardo Dobelin Barros
Stefan Diederichs
Lydia M. Duijvestijn
Michael Hammersley

Arindam Hazra
Corneliu Holban
Yan Li
Osai Osaigbovo
Andreas Plach
Ivan Portilla
Mukerji Saptarshi
Harinder P. Seera
Elisabeth Stahl
Clea Zolotow

ibm.com/redbooks

Redpaper

International Technical Support Organization

Performance and Capacity Implications for Big Data

January 2014

REDP-5070-00

Note: Before using this information and the product it supports, read the information in “Notices” on page v.

First Edition (January 2014)

This edition applies to IBM InfoSphere BigInsights and IBM InfoSphere Streams.

Contents

Notices v
Trademarks vi

Preface vii
Authors vii
Now you can become a published author, too! vii
Comments welcome viii
Stay connected to IBM Redbooks viii

Chapter 1. Introduction 1

Chapter 2. Big data overview and benefits 3
2.1 What “big data” means 4
2.2 Business needs that can benefit from big data solutions 4

Chapter 3. What big data means for performance and capacity 7
3.1 Volume 8
3.1.1 Scalability 8
3.1.2 The impact of big data on networking 9
3.1.3 Cloud services 10
3.2 Velocity 11
3.2.1 Access latencies 11
3.2.2 Rapid use and rapid data interpretation requirements 12
3.2.3 Response time 13
3.2.4 Impact of security on performance and capacity 14
3.3 Variety 15
3.3.1 Data types 15
3.3.2 Tuning 16
3.4 Veracity 20

Chapter 4. Performance and capacity for big data solutions today and tomorrow . . 21
4.1 Big data examples 22
4.1.1 Use scenarios 22
4.1.2 Client experiences 23
4.2 Emerging technologies for addressing big data challenges 24
4.2.1 Distributed processing 26
4.2.2 Streams computing 28
4.2.3 Indexing 29
4.2.4 Accelerating analytics for traditional databases 30
4.2.5 Big sorting 30
4.2.6 Packaged solutions 31

Chapter 5. Summary 33

Related publications 35
IBM Redbooks 35
Online resources 35
Help from IBM 36

Notices

This information was developed for products and services offered in the U.S.A.

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, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: 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 states 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 supply in any way it believes appropriate without incurring any obligation to you.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

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.

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 the names and addresses used by an actual business enterprise 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.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

BigInsights™
DB2®
developerWorks®
FlashSystem™
GPFS™
Guardium®
IBM®
IBM FlashSystem™
IBM PureData™
InfoSphere®
POWER7®
PowerLinux™
PureData™
Redbooks®
Redpaper™
Redbooks (logo) ®
Smarter Planet®
SPSS®
Symphony®
Velocity™

The following terms are trademarks of other companies:

Netezza, TwinFin, and N logo are trademarks or registered trademarks of IBM International Group B.V., an IBM Company.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

UNIX is a registered trademark of The Open Group in the United States and other countries.

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

Preface

Big data solutions enable us to change how we do business by exploiting previously unused sources of information in ways that were not possible just a few years ago. In IBM® Smarter Planet® terms, big data helps us to change the way that the world works.

The purpose of this IBM Redpaper™ publication is to consider the performance and capacity implications of big data solutions, which must be taken into account for them to be viable. This paper describes the benefits that big data approaches can provide. We then cover performance and capacity considerations for creating big data solutions. We conclude with what this means for big data solutions, both now and in the future.

Intended readers for this paper include decision-makers, consultants, and IT architects.

Authors

This paper was produced by a team of specialists from around the world working together with the International Technical Support Organization, Poughkeepsie Center.

Authors: Dave Jewell, Ricardo Dobelin Barros, Stefan Diederichs, Lydia M. Duijvestijn, Michael Hammersley, Arindam Hazra, Corneliu Holban, Yan Li, Osai Osaigbovo, Andreas Plach, Ivan Portilla, Mukerji Saptarshi, Harinder P. Seera, Elisabeth Stahl, Clea Zolotow

Editors: Renny Barrett, Matthew Gillard, John Shepherd, Robert T. Tatum

Reviewers: James Cristofero, Tom Farrell, Nin Lei, Melinda Mullett, Ji Deng Wang, Vicki Wojcik, Kun Yang

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

Comments welcome

Your comments are important to us!

We want our papers to be as helpful as possible. Send us your comments about this paper or other IBM Redbooks® publications in one of the following ways:

- Use the online **Contact us** form:

  ibm.com/redbooks

- Send your comments in an email message to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

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 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

1 Introduction

In a sense, the use of electronics to implement IT solutions has commonly been about big data since coming of age in the 20th century. Whether IT was used for compute-intensive applications or for implementing business processes, its value proposition has always been the ability to handle large amounts of data more quickly, more consistently, and more accurately than human beings.

This raises the question: Does big data represent an incremental change for IT or a major transformation? After all, technologies for data warehouses, data mining, and business intelligence have been with us for years, and manufacturing applications have long used analytics to respond quickly to variances found by sorting through large volumes of rapidly arriving process data.

The premise of this IBM Redpaper publication is that the proper answer is “both.” Just as cloud computing enables new ways for businesses to use IT because of many years of incremental progress in the area of virtualization, big data now enables new ways of doing business by bringing advances in analytics and management of both structured and unstructured data into mainstream solutions. Consider these examples:

- A major US retailer adds weather data to its distribution algorithms so that it can model delivery paths and can use disparate sources for improved logistics.
- A major Indian telecommunications firm analyzes billions of call records daily to target customers for special offers, which results in reducing churn and increasing loyalty among customers.
- A police department in a major US city installs traffic cameras throughout the city, and these traffic cameras can read license plates. The Department of Motor Vehicles uses this data to identify stolen vehicles in real time to get those cars off the street because many crimes are committed in stolen automobiles.

Big data solutions now enable us to change the way we do business, in ways that were not possible just a few years ago, by taking advantage of previously unused sources of information. In IBM Smarter Planet terms, big data helps us “change the way the world works.”

The focus of this paper is the performance and capacity implications of big data solutions, which must be taken into account for such solutions to be viable. This paper gives an overview of big data and the benefits that it offers, describes the performance and capacity aspects that must be considered in creating big data solutions, and suggests what this means for big data solutions, both now and in the future.

2

Big data overview and benefits

Down through the years of human history, the most successful decisions that were made in the world of business were based on the interpretation of available data. Every day, 2.5 quintillion bytes of data are created—so much that 90% of the data in the world today has been created in the last two years.[1] Correct analysis of the data is the key success factor in being able to make better decisions that are based on the data.

Given the quantity and complexity of the data that is being created, traditional database management tools and data processing applications simply cannot keep up, much less make sense of it all. The challenges for handling big data include capture, storage, search, sharing, transfer, analysis, and visualization. The trend to larger data sets is due to the additional information that can be derived from analysis of a single large set of related data, compared to separate smaller sets with the same total amount of data. Some estimates for the data growth are as high as 50 times by the year 2020.[2]

[1] "Apply new analytics tools to reveal new opportunities," IBM Smarter Planet website, Business Analytics page http://www.ibm.com/smarterplanet/us/en/business_analytics/article/it_business_intelligence.html

[2] John Gantz and David Reinsel, "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East." *IDC*, for EMC Corporation, December 2012 http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

2.1 What “big data” means

Big data is a phenomenon that is characterized by the rapid expansion of raw data. This data that is being collected and generated so quickly that it is inundating government and society. Therefore, it represents both a challenge and an opportunity. The challenge is related to how this volume of data is harnessed, and the opportunity is related to how the effectiveness of society’s institutions is enhanced by properly analyzing this information.

It is now commonplace to distinguish big data solutions from conventional IT solutions by considering the following four dimensions:

- *Volume.* Big data solutions must manage and process larger amounts of data.
- *Velocity.* Big data solutions must process more rapidly arriving data.
- *Variety.* Big data solutions must deal with more kinds of data, both structured and unstructured.
- *Veracity.* Big data solutions must validate the correctness of the large amount of rapidly arriving data.

As a result, big data solutions are characterized by real-time complex processing and data relationships, advanced analytics, and search capabilities. These solutions emphasize the flow of data, and they move analytics from the research labs into the core processes and functions of enterprises.

2.2 Business needs that can benefit from big data solutions

Big data is a technology to transform analysis of data-heavy workloads, but it is also a disruptive force. It is fueling the transformation of entire industries that require constant analysis of data to address daily business challenges. Big data is about broader use of existing data, integration of new sources of data, and analytics that delve deeper by using new tools in a more timely way to increase efficiency or to enable new business models. Today, big data is becoming a business imperative because it enables organizations to accomplish several objectives:

- Apply analytics beyond the traditional analytics use cases to support real-time decisions, anytime and anywhere
- Tap into all types of information that can be used in data-driven decision making
- Empower people in all roles to explore and analyze information and offer insights to others
- Optimize all types of decisions, whether they are made by individuals or are embedded in automated systems by using insights that are based on analytics
- Provide insights from all perspectives and time horizons, from historic reporting to real-time analysis, to predictive modeling
- Improve business outcomes and manage risk, now and in the future

In short, big data provides the capability for an organization to reshape itself into a contextual enterprise, an organization that dynamically adapts to the changing needs of its individual customers by using information from a wide range of sources.[3] Although it is true that many businesses use big data technologies to manage the growing capacity requirements of today's applications, the contextual enterprise[4] uses big data to enhance revenue streams by changing the way that it does business.

IBM consultants' experience with clients shows that big data use cases fall into these five major categories:

- Developing a 360-degree view of the customer
- Understanding operational analytics
- Addressing threats, fraud, and security
- Analyzing information that clients did not think was usable
- Offloading and augmenting data warehouses[5]

Each of these broad use case categories can lend itself to a different architecture and mix of technologies. Therefore, each calls for different priorities for performance and capacity.

[3] IBM Redpaper publication, *Smarter Analytics: Information Architecture for a New Era of Computing*, SG24-5012. http://www.redbooks.ibm.com/abstracts/redp5012.html

[4] Cheryl Wilson, "Making the Contextual Enterprise Possible with ODM," IBM Connections blog, 2013. https://www-304.ibm.com/connections/blogs/gooddecision/entry/making_the_contextual_enterprise_possible_with_odm?lang=en_us

[5] Doug Henschen, "IBM And Big Data Disruption: Insider's View." *Information Week*, 2013. http://ubm.io/1evdWay

3

What big data means for performance and capacity

This chapter presents a set of performance and capacity challenges that are associated with big data systems. It shows how these challenges might be different in nature, scale, or scope from other IT solutions.

A typical big data lifecycle involves the capture and input of data, managing and processing of the data, and the presentation of this data to the user. This chapter reviews the performance and capacity challenges from the perspective of the four big data dimensions, based on the lifecycle shown in Figure 3-1: volume, velocity, variety, and veracity.

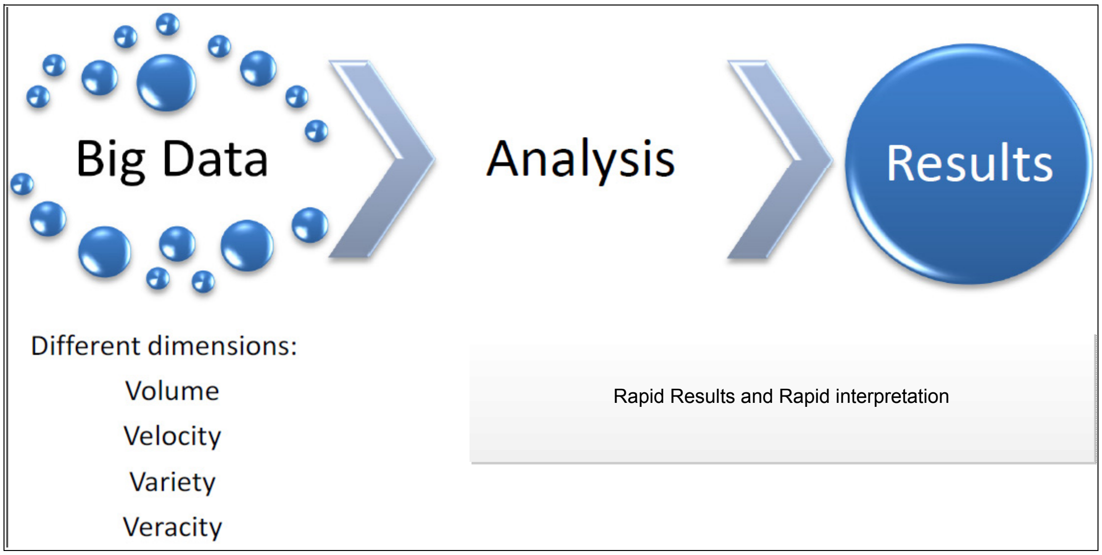


Figure 3-1 Lifecycle of big data

3.1 Volume

Many aspects of big data contribute to our interest in volume, and we describe a few of them in this section.

3.1.1 Scalability

The size of big data is easily recognized as an obvious challenge. Big data is pushing scalability in storage, with increases in data density on disks to match. The current Redundant Array of Independent Disks (RAID) approach that is in widespread use does not provide the level of performance and data durability that enterprises dealing with escalating volumes of data require. For example, committing data from memory to disk can increase overhead and cause processing delays if multiple disks are involved in each commit process.

Moreover, as the scale of data increases, the mean time between failures (MTBF) falls. For example, a system with a billion cores has an MTBF of one hour. The failure of a particular cluster node affects the overall calculation work of the large infrastructure that is required to process big data transactions.

Furthermore, a large percentage of the data might not be of interest. It can be filtered and compressed by an order of magnitude. The challenge is to filter intelligently without discarding data samples that might be relevant to the task. For example, data that is related to time or location might be subject to wide variances yet still be valid.

Data volume is increasing faster than computing resources and processor speeds that exist in the marketplace. Over the last five years, the evolution of processor technology largely stalled, and we no longer see a doubling of chip clock cycle frequency every 18 - 24 months.[1] Now, due to power constraints, clock speeds are largely stalled and processors are being built with increasing numbers of cores. In the past, people who were building large data processing systems had to worry about parallelism across nodes in a cluster. Now, you must deal with parallelism within a single node.

Unfortunately, parallel data processing techniques that were applied in the past for processing data across nodes do not directly apply for intra-node parallelism, because the architecture looks very different. For example, there are many more hardware resources, such as processor caches and processor memory channels, that are shared across cores in a single node. Furthermore, the move toward packing modern processors with multiple sockets (each with tens of cores) adds another level of complexity for intra-node parallelism.

Finally, with predictions of “dark silicon,” specifically that power considerations in the future are likely to prohibit us from using all of the hardware in the system continuously, data processing systems will probably be required to actively manage the power consumption of the processor. These unprecedented changes require us to rethink how data processing components are designed, built, and operated.

[1] Divyakant Agrawal, Philip Bernstein, Elisa Bertino, et al. “Challenges and Opportunities with Big Data.” Princeton University white paper, 2012. http://www.purdue.edu/discoverypark/cyber/assets/pdfs/BigDataWhitePaper.pdf

To resolve these problems, there are solutions such as *database sharding* (breaking data into small pieces and running the pieces in parallel). This approach is based on *shared nothing* architecture, which means no shared components between environments. Rather than storing application data in a single database on a single server with a shared processor, memory, and disk, the database is divided into several smaller shards, each of which can be hosted on independent servers with dedicated processor, memory, and disk. This greatly reduces resource contention.[2] The key benefit is that smaller databases are faster.

An increase in the number of nodes leads to a potential increase in the number of failures. Furthermore, there is a gradual shift from using hard disk drives (HDDs) to store persistent data. HDDs had far slower random I/O performance than sequential I/O performance. However, HDDs are increasingly being replaced by solid-state drives (SSDs), and other technologies, such as phase change memory, are around the corner. These new storage technologies require a rethinking of storage subsystem design for processing data, especially regarding high availability and fault tolerance.

In addition, the increase of unstructured data has a large impact on scalability. Data reliability relates to the limits of data density at tolerable device-level bit error rates. Traditional RAID does not provide the levels of data durability and performance for dealing with escalating volumes of data. For disks, there is the end of life, stress modes, and overheating in data centers to consider.

A shift that is underway is the transformative change of the traditional I/O subsystem. For many decades, HDDs were used to store persistent data. HDDs had far slower random I/O performance than sequential I/O performance. Data processing engines formatted their data and designed their query processing methods to work around this limitation.

Countering these problems requires new methods to improve rebuild times on high-density disk drives and to reduce susceptibility to data corruption induced by disk error. The ideal system maximizes the mean time to failure and minimizes the mean time to recovery. Even more important, it provides fault tolerance for a higher number of drive failures (that is, it minimizes the potential for concurrent failure and shrinks the exposure window).

Using erasure code algorithms that provide a faster protection level for bit error protection and flexibility also helps counter the problems.

3.1.2 The impact of big data on networking

In 2010, Eric Schmidt, then CEO of Google, was reported as saying: “Every two days, as much information is created as has been in existence since the dawn of civilization up until 2003.”[3] Big data invariably means that enterprises must handle larger amounts of data on existing network infrastructures. This presents a huge performance and capacity challenge, particularly for the use of Apache Hadoop as a building block for big data.

[2] Share Nothing, Shard Everything: dbShards Overview. codeFutures.com website, accessed 2014.
http://www.codefutures.com/dbshards

[3] M.G. Siegler, “Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003.” *TechCrunch*, 2010.
http://techcrunch.com/2010/08/04/schmidt-data/

A Cisco white paper explains the Hadoop data hierarchy this way:

> "The Hadoop Distributed File System (HDFS) is the first building block of a Hadoop cluster. To efficiently process massive amounts of data, it was important to move computing to where the data is, using a distributed file system rather than [a] central system for the data. A single large file is split into blocks, and the blocks are distributed among the nodes of the Hadoop cluster."[4]

An efficient and resilient network is a crucial part of a good Hadoop cluster. The nodes in a Hadoop cluster are interconnected through the network workload processing. A network is also crucial for writing data, reading data, signaling, and for operations of HDFS and the MapReduce infrastructure.

Therefore, the failure of a networking device affects multiple Hadoop data nodes. This means that a job might need to be restarted or more loads must be pushed to the available nodes, which makes jobs take a lot longer to finish. As a result, networks must be designed to provide redundancy with multiple paths between computing nodes and, furthermore, must be able to scale. Factors such as workload patterns, rack configurations within clusters, storage area network (SAN) access, and separation of data access networks from general networks need to be considered.

In addition, the network must be able to handle bursts effectively without dropping packets. For this reason, it is important to choose switches and routers with queuing and buffering strategies.[4]

There are a few other approaches that are commonly used to address the performance and capacity challenges:

- Use proprietary networks
- Design and segregate networks that are based on traffic (for example, separate a big data infrastructure management network from a data traffic network path)
- Apply data locality by taking an extract, *load*, and transform (ELT) approach (to process and analyze data where it is stored rather than using extract, *transform*, and load (ETL), which involves moving data twice)

Furthermore, the network design needs an acceptable and efficient *oversubscription ratio* (the advertised capacity versus the actual capacity). This handles situations where there is congestion at critical points in the network.

3.1.3 Cloud services

Big data and cloud services are two initiatives that are at the top of the agenda for many organizations. There is a view that cloud computing can provide the opportunity to enhance organizations' agility, enable efficiencies, and reduce costs. In many cases, cloud computing provides a flexible model for organizations to scale their big data capabilities, as evidenced by the inclusion of MapReduce in the offerings of Amazon Web Services. However, this needs to be done with careful planning, especially estimating the amount of data to analyze by using the big data capability in the cloud, because not all public or private cloud offerings are built to accommodate big data solutions.

In some cases, cloud computing environments can pose the following performance and capacity difficulties for big data:

[4] "Big Data in the Enterprise: Network Design Considerations." Cisco white paper, 2011. http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9670/white_paper_c11-690561.pdf

- The movement of large data sets into and out of the cloud can be affected by the degradation in WAN transfer speeds that occurs over long distances when you are using traditional transfer protocols. It can also be affected by the “last foot” bottleneck inside of the cloud data center, which is caused by the HTTP interfaces with the underlying object-based cloud storage.
- Where there is a need to analyze low-latency, real-time data, you might need to use different approaches. If you do not have the performance necessary to process real-time data without introducing latency, it might make the results too stale to be useful.[5]
- In other cases, the use of cloud technologies might not be appropriate for use with big data analysis, because it is more suitable for variable and random use among users. Big data analysis requires a dedicated infrastructure that is used at full capacity for hours at a time. That is, the analysis is normally performed by batch jobs. This is not to say that the cloud cannot be used for storage; however, it requires careful design to cater to big data analysis.

3.2 Velocity

This section describes the performance and capacity challenges that result from the increased speed of the flow of data through organizations. Some of the challenges that result from the increased velocity of big data include access latencies, the need for rapid use of the data, the need for faster response times, and the impact on the organization’s security mechanisms.

3.2.1 Access latencies

Access latencies create bottlenecks in systems in general, but especially with big data. The speed at which data can be accessed while in memory, network latency, and the access time for hard disks all have performance and capacity implications. For big data, data movement is usually not feasible, because it puts an unbearable load on the network. For example, moving petabytes of data across a network in a one-to-one or one-to-many fashion requires an extremely high-bandwidth, low-latency network infrastructure for efficient communication between computer nodes.

Consider the performance and capacity implications of Hadoop processing on disk systems that are not designed for the speed of data movement that big data requires. Hadoop is built on top of the distributed file system called Hadoop Distributed File System (HDFS). Usually, an HDFS writes the data three times for a triple mirroring scheme (replication). This is to ensure that no data is lost, because the system was originally designed for “just a bunch of disks” (JBOD), not for enterprise-class disk systems. This ensures that the Hadoop cluster can be scaled at a low cost compared to enterprise disk systems. Enterprise disk systems, such as SANs, are typically avoided because of the high level of data traffic that needs to be sustained as the cluster scales upward. For example, given a 5-node cluster with a 50 MB/s throughput on each node, the total amount of throughput that is required is 250 MB/s. If the cluster is scaled to 10 nodes, the total throughput increases.

Because there is a triple redundancy feature, there are capacity implications for disk capacity. Furthermore, the replication can create loads on the servers (processor, memory, and I/O), because these are not off-loaded data replication processes. This can cause an immense load on the network while the systems try to handle the traffic.

[5] Seth Payne, “How Cloud Computing Democratizes Big Data.” Readwrite.com, 2013 http://readwrite.com/2013/06/07/how-cloud-computing-democratizes-big-data#awesm=~oiudiSdGtPlwpJ

There is another impact to consider with potential disk failures. If they occur, there can be periods of high I/O, lasting hours to days, that result when the system must rebalance itself to ensure that each node is replicated the correct number of times. This is the performance and capacity load of setting up the systems.

Big data uses different types of analytics, such as “adaptive predictive models, automated decision making, network analytics, analytics on data-in-motion, and new visualization.”[6] Previously, data was pre-cleaned and stored in a data mart. Now, most or even all source data is retained. Furthermore, new types of feeds, such as video or social media feeds are available (Twitter, for instance). Addressing all of these feeds has a computational cost that pushes query use throughout hardware components for server, I/O, memory, network, and SAN to higher-than-expected levels. Traditional performance and capacity techniques can be used in combination with newer big data-specific analytics techniques to ensure optimal processing time of analytic queries in a big data system.

3.2.2 Rapid use and rapid data interpretation requirements

According to *The Economist*, “Data is becoming the new raw material of business: an economic input almost on a par with capital and labor.”[7] It is crucial in today’s fast-paced business climate to derive rapid insight from data. Consequently, agility is essential for businesses. Successfully taking advantage of the value of big data requires experimentation and exploration, and both need to be done rapidly and in a timely manner.[8]

All of these requirements result in performance challenges as the amount of data that moves into the system increases. First, there is the challenge of whether there is enough I/O and network bandwidth when you are pushing the data to storage. Second, there is the challenge of checkpointing this data with regard to in-memory analytics systems. These systems perform computations in memory at a certain point to ensure that the results of a particular step of a complex calculation are not lost. That data needs to be written to a local disk. While data is being written to disk for safekeeping (that is, checkpointing), the node is not performing computations. As the amount of data that needs to have a checkpoint increases, the amount of time that is required for the process to complete increases. To reduce the time that checkpointing takes and make more time available for computations, you need more disks, or a faster, lower-latency disk (such as solid-state drives), or flash memory.

The size of big data invariably means that it takes longer to identify potential bottlenecks that might affect the performance of the system. Therefore, the system must be designed to respond quickly. Cluster architectures with fast, proprietary, low-latency networks and large memory (such as RAM) are typical approaches to ensure rapid response from big data systems. To facilitate speed and real-time results, a new approach is emerging in which subsets of the big data are held and processed within a server’s fast local memory rather than having to access the same information from slow disk storage or from another server. Such *in-memory computing* enables dramatically faster computation and analysis. In the world of finance, where speed is everything, the use of in-memory computing can help increase a firm’s competitive advantage.[9]

[6] John Hagerty and Tina Groves, “Unlock Big Value in Big Data with Analytics: An IBM Redbooks Point-of-View publication.” *IBM Redbooks* publications, 2013.
http://www.redbooks.ibm.com/abstracts/redp5026.html

[7] Special report. “Data, data everywhere.” *The Economist*, 2010.
http://www.economist.com/node/15557443

[8] Edd Dumbill. “What is big data? An introduction to the big data landscape.” O’Reilly, 2012.
http://strata.oreilly.com/2012/01/what-is-big-data.html

[9] Scott Sellers. “Overcoming the big performance challenge of Big Data in finance.” *Computerworld* blog, 2012.
http://blogs.computerworld.com/20084/overcoming_the_big_performance_challenge_of_big_data_in_finance

Another point to consider is that for typical cluster architectures, all of the infrastructure (for example, computing nodes) needs to be of the same specification in terms of memory, processors, I/O bandwidth, and operating system version. The cluster runs at the speed of the node with the lowest specification. It is important to remember this as the infrastructure scales to meet performance needs.

3.2.3 Response time

Response times for results are still critical, despite the increase of data size. To ensure speed and real-time feedback from big data, a new approach is emerging where data sets are processed entirely within a server's memory.

Several vendor solutions offer in-memory analytics to address the requirements for real-time analysis results (for example, IBM InfoSphere® Streams and BigInsights™, SAS Visual Analytics, and SAP HANA). These apply in scenarios where there is a high velocity of data or real-time return of results. The main limitation with traditional business intelligence technologies is the time that it takes to read from and store to disk (disk I/O). By storing the data in memory (RAM), this limitation is removed. However, this needs to be balanced with the higher cost of RAM, compared to disk storage, to determine whether business requirements justify the additional expense.[10]

Typically, in-memory technologies have Java virtual machines (JVMs) running on the application tier. JVMs use garbage collection to reclaim heap space from objects that are no longer being used or are out of reach. They "pause" current JVM processing to perform this task. When the size of the JVM memory heap is large (64-bit JVM),[11] the garbage collection behaves in unpredictable ways and can cause large pauses.[12] That is not suitable for high-velocity data analysis or real-time return of results. However, there are tuning techniques that can minimize this problem, such as the use of compressed references.[13]

As more companies move toward big data and test the limits of applications such as the Hadoop cluster, the JVM issues are becoming more apparent. Different solutions are being proposed or made available to address them. For example, a new version of Hadoop implements CRC32C by using hardware support to improve performance.[14]

People who have the skill set required for tuning large JVMs are rare. Therefore, it is advisable to engage the primary vendors who are deploying the in-memory analytic technology to configure the JVM. This helps to ensure that it runs in an optimal manner, rather than taking the risk of doing it in-house and then encountering difficulty in resolving performance and capacity problems.

[10] This paragraph and the next are summarized from these sources:

Scott Sellers, "Overcoming the big performance challenge of Big Data in finance." *Computerworld*, 2012
http://blogs.computerworld.com/20084/overcoming_the_big_performance_challenge_of_big_data_in_finance

Scot Petersen, Editor, "In-Memory Analytics and Big Data: A Potent Mix?" *TechTarget*, 2013
http://docs.media.bitpipe.com/io_10x/io_109628/item_682431/In-Memory_Analytics_and_Big_Data_A_Potent_Mix_hb_final.pdf

[11] Nikita Salnikov-tarnovski, "Should I use a 32- or a 64-bit JVM?" DZone.com, 2012.
http://java.dzone.com/articles/should-i-use-32-or-64-bit-jvm

[12] Frequently Asked Questions About the Java HotSpot VM. Oracle.com.
http://www.oracle.com/technetwork/java/hotspotfaq-138619.html#64bit_heap

[13] Tomas Nilsson, "Understanding Compressed References." The JRockit Blog, Oracle.com, 2010.
https://blogs.oracle.com/jrockit/entry/understanding_compressed_refer

[14] Hadoop Common, Hadoop-7446, Implement CRC32C native code using SSE4.2 instructions. Apache.com, accessed 2014.
https://issues.apache.org/jira/browse/HADOOP-7446

Due to the practical limitations of garbage collection (GC) with large memory for a big data environment, multiple solutions have been proposed. A few examples follow:

- Terracotta BigMemory[15] uses a region of memory outside of the JVM heap for cached data, which it calls an "off-heap store." This heap is not subject to GC. Therefore, for temporary data, an appropriate heap can be allocated by the JVM. This allows the garbage collector to work on a smaller heap size.
- The HotSpot VM has an extension that is called "GC Invisible Heap,"[16] which acts as a second-level cache and works with ParNew+CMS.

  Note: These are two GC types:

  - Parallel New works with the young (recent) data generation.
  - Concurrent Mark-Sweep works with the older (longer-lasting) data.
- A new G1 GC type collector[17] is optimized for multiprocessor machines with large memories and is fully supported in JDK7 Update 4 and later releases.
- Azul Continuously Concurrent Compacting Collector (C4) garbage collector[18] supports simultaneous-generational concurrency.[19] During long periods of concurrent full heap collection, this GC technology continuously collects concurrent young generations. This avoids performing a stop-the-world[20] GC operation or sacrificing response time.

3.2.4 Impact of security on performance and capacity

A further consideration is the impact of big data velocity on the manner in which security mechanisms are applied to existing data sets. The increased velocity of data corresponds to an increase in security-relevant data. According to Tim Mather of KPMG, "Many big data systems were not designed with security in mind."[21] Typically, a high level of security is not needed for all of an organization's data. However, even if the amount that needs to be secured is only 5%, that amount is still huge from a big data perspective and has an impact on performance and capacity requirements for the underlying systems.

The security mechanisms need to be applied in a manner that does not increase access latency. In addition, big data technology enables massive data aggregation beyond what was previously possible. Therefore, organizations need to make data security and privacy high priorities as they collect more data in trying to get a single view of the customer.

[15] "BigMemory: Max or Go? Get the version that's right for you." Terracotta.org, accessed 2014. http://terracotta.org/products/bigmemory

[16] Kris Mok, "JVM @ Taobao," presentation at QCon Hangzhou 2011. Slideshare.net. http://www.slideshare.net/RednaxelaFX/jvm-taobao

[17] Java HotSpot Garbage Collection page, Oracle.com, accessed 2014. http://www.oracle.com/technetwork/java/javase/tech/g1-intro-jsp-135488.html

[18] Gil Tene, Balaji Iyengar, Michael Wolf, "C4: The Continuously Concurrent Compacting Collector" white paper. Azul Systems, Inc., not dated. http://www.azulsystems.com/products/zing/c4-java-garbage-collector-wp

[19] The different generations are collected at the same time, while processing continues.

[20] Processing stops while garbage is collected.

[21] Michael Cooper and Peter Mell, "Tackling Big Data" slide presentation. NIST Information Technology Laboratory, Computer Security Division, US Department of Commerce, not dated. http://csrc.nist.gov/groups/SMA/forum/documents/june2012presentations/fcsm_june2012_cooper_mell.pdf

3.3 Variety

Big data encompasses various data types: structured, unstructured, and semi-structured. All of these need to be harnessed for an organization to get the true value of big data. Integrating and analyzing various data sources enable organizations to gain great insight that is derived from the data that is available for decision making. Businesses that broaden the range of data sources also achieve better value.[22] This involves combining a wide range of data formats that are traditionally not stored on the same platform and are constantly changing. These unique circumstances result in several performance and capacity challenges.[23] Most big data projects face more challenges from variety and fewer from data volumes.[24]

3.3.1 Data types

One of the crucial challenges that affect performance and capacity in a big data system arises from the variety of data types that can be introduced during typical processing cycles. These challenges can arise for these reasons:

- Growth, necessitating the addition of new systems, which can result in an uncontrolled heterogeneous landscape in the enterprise (such as a plethora of types of systems)
- The introduction of new systems that provide data but introduce challenges in identifying its relevance in big data systems

To solve the impact on the performance and capacity of big data systems, there are several approaches that organizations can take:

- Define a comprehensive taxonomy that aids in the design and increases usability of the big data system.
- Identify the numerous ways that data objects of the same type are represented across the heterogeneous enterprise.
- Constantly review data changes and control new nomenclature referring to the same data element.[25]

Much of big data is unstructured. By nature, it is fast to write but takes more time to read and use. One approach to solving this problem is the use of flash memory, such as IBM FlashSystem™,[26] which can accelerate retrieval of unstructured data because of flash memory's lower access latency.

[22] Press release, "Gartner Survey Finds 42 Percent of IT Leaders Have Invested in Big Data or Plan to Do So Within a Year." Gartner, Inc.
http://www.gartner.com/newsroom/id/2366515

[23] Mahesh Kumar, "Today's Big Data Challenge Stems From Variety, Not Volume or Velocity." Janalta Interactive Inc., 2012.
http://www.techopedia.com/2/29109/trends/big-data/todays-big-data-challenge-stems-from-variety-not-volume-or-velocity

[24] "Taming Data Variety and Volatility is Key for Big Data Analytics." Speaking of Analytics, The Lavastorm Blog, not dated.
http://www.lavastorm.com/blog/post/taming-data-variety-and-volatility-is-key-for-big-data-analytics/

[25] Mahesh Kumar, "Today's Big Data Challenge Stems From Variety, Not Volume or Velocity." Techopedia.com, 2012.
http://www.techopedia.com/2/29109/trends/big-data/todays-big-data-challenge-stems-from-variety-not-volume-or-velocity

[26] Flash storage and solutions, IBM System Storage web page.
http://www.ibm.com/systems/storage/flash/

3.3.2 Tuning

The rise of information from a variety of sources, such as social media, sensors, mobile devices, videos, and chats, results in an explosion of the volume of data. Previously, companies often discarded the data because of the cost of storing it. However, with frameworks such as Hadoop and relatively inexpensive commodity servers, it is now feasible for companies to store the data.[27] Along with inexpensive servers and storage comes the potential complexity of deployment and management of a very large infrastructure. Similarly, tuning this infrastructure for the expected big data workloads is still a challenge for which proven, reusable patterns are still in the early stages of development.

Due to the popularity of Hadoop[28] for big data, the rest of the section focuses on Hadoop. It describes different monitoring tools that aid in tuning Hadoop clusters and presents an approach to tuning them. This approach can be used as a starting point and enhanced as relevant to the application.

Hadoop is a well-known open source framework for big data distributed applications.[29] Tuning a Hadoop cluster requires understanding the Hadoop framework and all of the components of the stack, which include Hadoop MapReduce[30], the JVM, the network, OS, hardware, background activities, and, possibly, the BIOS,[31] as shown in Figure 3-2.

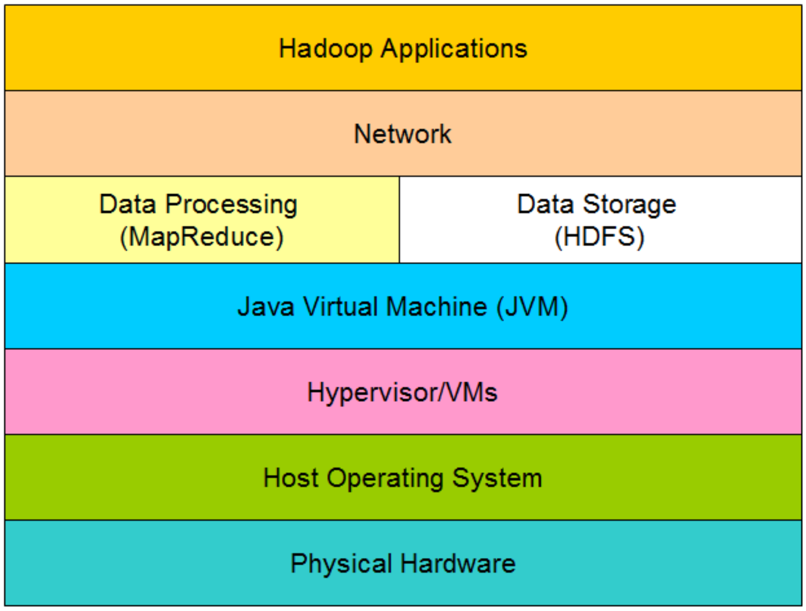


Figure 3-2 Technology stack of a Hadoop environment

[27] Brian Proffitt, “The Real Reason Hadoop Is Such A Big Deal In Big Data.” ReadWrite.com, 2013. http://readwrite.com/2013/05/29/the-real-reason-hadoop-is-such-a-big-deal-in-big-data#awesm=~oaN48yeiCjoBMy

[28] Hortonworks blog. The rise of Hadoop’s dominance in the big data market. Hortonworks.com, 2013. http://hortonworks.com/blog/the-rise-of-hadoops-dominance-in-the-big-data-market

[29] Hadoop page. Apache.org website, accessed 2014. http://hadoop.apache.org/

[30] MapReduce is a programming method that allocates the processing of a large data set across multiple nodes and then combines the results.

[31] Shrinivas Joshi, “Apache Hadoop Performance-Tuning Methodologies and Best Practices” (PDF from Advanced Micro Devices, Inc.). *Admin Magazine*, 2011. http://www.admin-magazine.com/HPC/Vendors/AMD/Apache-Hadoop-Performance-Tuning-Methodologies-and-Best-Practices

The Hadoop framework has more than 150 configurable parameters that can affect performance. A comparable number of tunable parameters exist for each layer in the rest of the stack. Therefore, the scope of the tuning exercise spans the entire environment. These are some of the Hadoop configurable parameters that can affect performance:

- **`dfs.blocksize:`** "The default block size for new files, in bytes."[32] The default size is 128 MB. The value of dfs.blocksize is dependent on cluster and data set size. A small value can cause extra overhead in creating too many map tasks. Therefore, cluster and data set size must be taken into account when you set the value.
- **`mapred.tasktracker.reduce.tasks.maximum:`** "The maximum number of **`reduce`** tasks that are run simultaneously by a task tracker."[33] This parameter and the parameter for mapred.tasktracker.map.tasks.maximum affect processor use, because increasing these values appropriately increases multitasking improvement throughput.[34]
- **`mapred.tasktracker.map.tasks.maximum:`** "The maximum number of map tasks that are run simultaneously by a task tracker."[33] This parameter and mapred.tasktracker.reduce.tasks.maximum also affect processor use, because increasing these values appropriately increases multitasking to improve throughput.[34]
- **`io.sort.mb:`** "The total amount of buffer memory to use while sorting files, in megabytes."[33] This parameter affects the I/O times, because the greater the value, the fewer spills to the disk.[34]
- **`mapred.min.split.size:`** "The minimum size of a chunk that map input needs to be split into, in bytes."[33] This parameter can cause some data blocks not to be local to the map tasks that are handling them if the value of the split is larger than the block size.[35] Therefore, block size and **`maximumsplit.size`** parameters also need to be examined before any changes.
- **`io.sort.factor:`** "The number of streams to merge simultaneously while sorting files. This parameter affects I/O time for map/reduce tasks.[34]

Having appropriate monitoring tools is critical to knowing the overall health of the environment and for aiding in tuning. Table 3-1 on page 18 presents some of the commercial and open source tools that are available to monitor Hadoop clusters.

[32] Hadoop project descriptions (table), Apache.org website. `http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/hdfs-default.xml`

[33] Hadoop parameters descriptions (table), Apache.org website. `http://hadoop.apache.org/docs/r1.0.4/mapred-default.html`

[34] Yu Li, "Analyze and optimize cloud cluster performance." IBM developerWorks, 2011. `http://www.ibm.com/developerworks/cloud/library/cl-cloudclusterperformance/`

[35] Tom White. *Hadoop The Definitive Guide*, First Edition. O'Reilly Media, 2009.

Table 3-1 Monitoring and profiling tools

Tools	Description
Ganglia[a]	Provides an overview of cluster use Current Hadoop version comes with the Ganglia plug-in to monitor Hadoop-specific metrics
Nagios[b]	Infrastructure monitoring tool with alert facility
Splunk[c]	Provides a way to collect, analyze, and visualize machine data Specific Hadoop software is available, such as Hunk and Splunk Hadoop Connect
Starfish[d]	Log analyzer
White Elephant[e]	Log aggregator and dashboard for Hadoop cluster visualization
Apache Vaidya[f]	Tool to diagnose map/reduce jobs performance
Dstat, vmstat, top, iostat, sar, netstat, OProfile (and more)	Standard Linux OS monitoring utilities
Hprof[g]	Java heap and processor profiling
JConsole,[h] VisualVM[i]	JVM instrumentation tools
strace, perf	Linux profiling tools

a. http://ganglia.sourceforge.net
b. http://www.nagios.org
c. http://www.splunk.com
d. https://www.cs.duke.edu/starfish
e. https://github.com/linkedin/white-elephant
f. http://hadoop.apache.org/docs/stable/vaidya.html
g. http://docs.oracle.com/javase/7/docs/technotes/samples/hprof.html
h. http://docs.oracle.com/javase/6/docs/technotes/tools/share/jconsole.html
i. http://docs.oracle.com/javase/6/docs/technotes/guides/visualvm/index.html

Other applications that allow monitoring of Hadoop clusters are Horton Ambari[36] and Cloudera Manager.[37]

Figure 3-3 highlights the steps to follow as part of a typical Hadoop cluster-tuning activity. For successful completion, other activities, such as the following tasks, are necessary before proceeding with tuning:

- Defining the performance objectives and associated metrics
- Identifying the expected workload to be handled on the Hadoop cluster
- Identifying whether to use a test or a production environment for the tuning activity

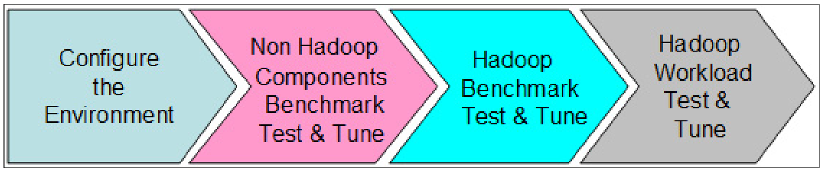

Configure the Environment
Non Hadoop Components Benchmark Test & Tune
Hadoop Benchmark Test & Tune
Hadoop Workload Test & Tune

Figure 3-3 Tuning steps

Tip: Conduct the benchmarking and tuning activities before the cluster is put into service.

[36] Apache Ambari page, Hortonworks website, accessed 2014. http://hortonworks.com/hadoop/ambari

[37] Cloudera Manager: End-to-End Administration for Hadoop. Cloudera.com website, accessed 2014. http://www.cloudera.com/content/cloudera/en/products-and-services/cloudera-manager.html

However, individual activities can be conducted on their own or as a whole, depending on the requirement and availability of necessary resources. For example, you can conduct the non-Hadoop component benchmark testing and tuning when the servers or cluster are made available (commissioned). Similarly, the Hadoop benchmark test and tune step can be performed after the installation of the Hadoop framework on the cluster. Alternatively, you can perform these steps after the environment is set up and available. Table 3-2 presents the details of each tuning step that shown in Figure 2.

Table 3-2 Description of tuning steps

Steps	Description	Tasks
1	Configuring the environment	► Verify that hardware is correctly configured. ► Verify that software components are up-to-date with correct patches (Linux distribution, JVM, Hadoop distribution). ► Install monitoring tools for instrumentation of the environment. Verify that they work and are capturing the correct information.
2	Non-Hadoop components benchmark test and tune	► Stress test each non-Hadoop component in the cluster. For example, the Netperf [a] benchmark can be used to measure network performance. Similarly, the STREAM [b] benchmark is used to verify memory performance. ► If required, diagnose bottlenecks and tune the components. Retest to validate the fixes.
3	Hadoop benchmark test and tune	► Stress test the Hadoop framework setup by running Hadoop benchmarks, such as TeraSort, TestDFSIO, MRBench, or NNBench. These benchmarks come with Hadoop distributions. ► If a benchmark fails, diagnose the bottleneck and tune the framework. ► Retest the framework to form a baseline for the real Hadoop workload.
4	Hadoop workload test and tune	► Run the expected production workload against the Hadoop cluster. ► Analyze the results and compare them to the acceptable metrics that are captured as part of requirements gathering process. ► If the test fails, diagnose and tune the environment. ► Retest to verify that the bottleneck is fixed.

[a] http://www.netperf.org/netperf/NetperfPage.html

[b] https://www.cs.virginia.edu/stream/

Note: Stress tests are run to evaluate the component and infrastructure behavior when it is pushed beyond the expected workload conditions.

3.4 Veracity

Every day, 2.5 quintillion bytes of data are created. This data comes from digital pictures, videos, posts to social media sites, intelligent sensors, purchase transaction records, and cell phone GPS signals, to name a few.[38] In short, big data is messy data.[39] Veracity deals with uncertain or imprecise data.[40] If the data is error-prone, the information that is derived from it is unreliable, and users lose confidence in the output. Cleaning the existing data and putting processes in place to reduce the accumulation of dirty data is crucial.[41]

It is very important that companies get data quality right.[42] To achieve improved efficiency, they need better-quality data.

To address the performance and capacity challenges that arise from lack of veracity, it is important to have data quality strategies and tools as part of a big data infrastructure. The aim of the data quality strategies is to ascertain “fit for purpose.” This involves evaluating the intended use of big data within the organization and determining how accurate the data needs to be to meet the business goal of the particular use case. The data quality approaches that the organization adopts need to include several strategies:

- Definition of data quality benchmarks and criteria
- Identification of key data quality attributes (such as timeliness and completeness)
- Data lifecycle management and compliance
- Metadata requirements and management
- Data element classification[43]

Furthermore, big data governance is critical. The approach needs to be collaborative and to be clear about what needs to be 100% precise and what is “good enough,” in percentage terms. The collaborative process must identify the strategic data stores and the specific data items that are critical from a governance perspective.[44]

Security also plays a key role by ensuring that fake, fraudulent, or unauthorized data is not introduced into big data. Data must be properly protected and not distributed to unauthorized recipients.

[38] Roberto V. Zicari, “Big Data: Challenges and Opportunities.” Goethe University Frankfurt presentation, no date. http://gotocon.com/dl/goto-aar-2012/slides/RobertoV.Zicari_BigDataChallengesAndOpportunities.pdf

[39] Loraine Lawson, “Getting Big Data Ready for Business: Tackling Big Data Governance and Quality.” *IT Business Edge*, 2013. http://www.itbusinessedge.com/blogs/integration/getting-big-data-ready-for-business-tackling-big-data-governance-and-quality.html

[40] Dwaine Snow, “Adding a 4th V to BIG Data - Veracity.” Blog: Dwaine Snow’s Thoughts on Databases and Data Management, 2012. http://dsnowondb2.blogspot.com/2012/07/adding-4th-v-to-big-data-veracity.html

[41] Jason Tee, “Handling the four V’s of big data: volume, velocity, variety, and veracity.” eServerSide.com, 2013. http://www.theserverside.com/feature/Handling-the-four-Vs-of-big-data-volume-velocity-variety-and-veracity

[42] Kathleen Hall, “Data quality more important than fixating over big data, says Shell VP.” ComputerWeekly.com, 2013. http://www.computerweekly.com/news/2240186887/Data-quality-more-important-than-fixating-over-big-data-says-Shell-VP

[43] Virginia Prevosto and Peter Marotta, “Does Big Data Need Bigger Data Quality and Data Management?” *Verisk Review*, not dated. http://www.verisk.com/Verisk-Review/Articles/Does-Big-Data-Need-Bigger-Data-Quality-and-Data-Management.html

[44] Loraine Lawson, “Getting Big Data Ready for Business: Tackling Big Data Governance and Quality.” *IT Business Edge*, 2013. http://www.itbusinessedge.com/blogs/integration/getting-big-data-ready-for-business-tackling-big-data-governance-and-quality.html

4

Performance and capacity for big data solutions today and tomorrow

Big data solutions are typically based on a hardware and software platform that covers big data key points, such as integration, analysis, visualization, development, workload optimization, security, and governance.

There are various big data uses where performance and capacity challenges are addressed by using emerging technologies that are provided by the big data platform.

4.1 Big data examples

This section presents use scenarios and client experiences.

4.1.1 Use scenarios

Table 4-1 shows performance and capacity use scenarios for big data solutions are, in general, based on big data's dimensions, use cases, capabilities, and entry points. When we consider the four dimensions of big data, five ways to put it to use come to mind (two of them are in the fourth dimension).

Table 4-1 Big data use cases by dimension

Big data			
Dimensions	**Use cases**	**Capabilities**	**Entry points**
1. Volume	1. Data warehouse augmentation (Integrate big data and data warehouse capabilities to increase operational efficiency.)	1. Data warehousing (Deliver deep operational insight with advanced in-database analytics.)	1. Simplify warehouse (IBM Netezza®)
2. Velocity	2. Big data exploration (Find, visualize, understand all big data to improve business knowledge.) 3. Enhanced 360^{o} View (Achieve a true unified view, incorporating internal and external sources.)	2. Stream computing (Drive continuous analysis of massive volumes of streaming data with submillisecond response times.)	2. Unlock big data (IBM InfoSphere Data Explorer) 3. Analyze streaming data (IBM InfoSphere Streams)
3. Variety	4. Operations analysis (Analyze various machine and operational data for improved business results.)	3. Apache Hadoop-based analytics (Process and analyze any data type across commodity server clusters.)	4. Reduce Cost with Hadoop (IBM BigInsights) 5. Analyze Raw Data (IBM BigInsights™)
4. Veracity	3. Enhanced 360^{o} view (Achieve a true unified view, incorporating internal and external sources—indirectly, the master data management, or MDM, concept.) 5. Security and intelligence extension (Lower risk, detect fraud and monitor cyber security in real time.)	2. Stream computing (Monitor for "truth.")	3. Analyze streaming data (IBM Streams, IBM Guardium®). 5. Analyze raw data (IBM BigInsights, IBM Guardium)

The main differences between big data use cases and traditional data warehouse or business intelligence (BI) applications are the nature and speed of the data under consideration. Typically, big data applications are thousands of times larger and require faster response time than traditional BI applications.

4.1.2 Client experiences

The examples that follow are just a few of the many performance-related and capacity-related examples of client experiences with big data solutions.

Wind power company

A wind power company optimizes capital investment decisions that are based on petabytes[1] (PB) of information:

- Big data technology used: Distributed processing.
- Need: Model the weather to optimize placement of turbines to maximize power generation and longevity.
- Benefits: Reduces time that is required to identify placement of turbines from weeks to hours. Reduces IT footprint and costs and decreases energy consumption by 40%, while increasing computational power.
- Performance and capacity improvements: Incorporates 2.5 PB of structured and semi-structured information flows. Data volume is expected to grow to 6 PB.

Internet company

An Internet provider uses streaming to gain customer insight and increase revenue:

- Big data technology used: Stream computing.
- Need: Required a system capable of processing 5,000 transactions per second, 18 million transactions per hour.
- Benefits: Increased subscribers by 100% due to pay-as-you-go payment plan flexibility and convenience.
- Performance and capacity improvements: Provides insight into subscriber demands and content use. Increased individual subscriber use of both pay-as-you-go and flat-rate services. Enables subscribers to monitor their Internet use and upgrade to the flat-rate plan if their use is higher than the maximum of 1000 MB per month.

Media company

A media company gains insights from data warehouse systems and social media:

- Big data technology used: IBM PureData™ Systems for Analytics.
- Need: Client needed to work with an extremely low level of data granularity and run these analytics at the speed of thought. Client wants to gain insights from data warehouse systems and social media to understand how its consumers make purchasing decisions.
- Benefits: Gained operational efficiencies and reduced the size of their database administration team due to speed, simplicity, and performance.
- Performance and capacity improvements: Provides the ability to deliver on-demand analytics. Compared to the earlier TwinFin® architecture, the PureData System performed 2x to 3x better on batch processes and anywhere from 3x to 10x better on the concurrency workload.

[1] 1024 terabytes or 1 quadrillion bytes

Global aerospace manufacturer

A global aerospace manufacturer increases its staff's access to critical information:

- Big data technology used: Data exploration.
- Need: Improve operational efficiencies by providing a unified search, discovery, and navigation capability to provide fast access to relevant information across the enterprise.
- Benefits: Placed 50 more aircraft into service worldwide during the first year without a staffing increase. Saved USD36 million per year in supporting the 24x7 aircraft-on-ground program.
- Performance and capacity improvements: Provides supply chain insight to reduce cycle time, which results in saving millions of dollars on critical parts deliveries.

Large retailer

A large retailer with increasing data satisfies batch window performance:

- Big data technology used: Application that uses product codes. (Variability of the workload was a considerable influence because of the peaks in replenishing stock during certain periods.)
- Need: Client was facing a projection of enormous growth in their volume of data.
- Benefits: Meeting the Service Level Agreement batch window, which is their most important *key performance indicator* (KPI).
- Performance and capacity improvements: Tuned applications, added flash memory, and increased network Gigabit Ethernet to eliminate bottlenecks. These drove the system infrastructure, which resulted in meeting their KPI objectives.

4.2 Emerging technologies for addressing big data challenges

Table 4-2 on page 25 presents emerging technologies that are beginning to address big data platform challenges.

Table 4-2 Summary of big data emerging technologies

Areas of innovation	Product, tool, or service examples	Supporting tools, components, and characteristics	Specific innovations, including those relevant to performance and capacity
Distributed processing	Hadoop: Supports batch-oriented read-intensive applications. Able to distribute and manage data across many nodes and disks	► Google MapReduce ► Apache Hadoop Distributed File System (HDFS) ► Supporting open source tools ecosystem[a]	► Accelerated input and aggregation processing ► Distributed data to use redundancy and parallelism
	IBM BigInsights: Analytical platform for processing large volumes of persistent "big data at rest" in a highly distributed and efficient manner	► Hadoop + open source tools ecosystem ► Big SQL[b] ► Adaptive MapReduce ► IBM General Parallel File System (GPFS™) File Placement Optimizer	► Load from IBM DB2®, Netezza, Teradata ► Integration with Netezza, IBM DB2 for Linux, UNIX, and Windows with DPF, R Statistics ► Platform enhancements ► Text analytics
Streams processing	InfoSphere Streams technology	► Continuous ingestion and analysis ► Applications are partitioned into software components ► Components are distributed across stream-connected hardware hosts	Accelerated Java processing Optimized for multiple data types and analytical tasks: ► IBM SPSS® statistical scoring ► Advanced text analytics ► Complex event processing
Search, discovery, and navigation	IBM InfoSphere Data Explorer: Foundation for search, discovery, and navigation systems. Intended to optimize performance, storage use, scalability, and federation of large amounts of data from multiple sources.		► Position-based indexing ► Index distribution and replication
Accelerated analytics for traditional databases	IBM DB2 with BLU Acceleration: Speeds analytics and reporting by using dynamic in-memory columnar technologies and other performance enhancements		► In-memory processing ► Columnar processing ► Actionable compression[c] ► Parallel vector processing ► Data "skipping"
Big sorting	IBM Terasort Benchmark Optimization Practice: Applies terasort benchmark (sorting one terabyte of data) to help clients optimize their big data solutions in a balanced manner		► Correlation-based analysis tool identifies bottlenecks ► Empirical rules for diagnosing and optimizing performance
Packaged solutions	IBM analytical DBMS: MPP Data Warehouse appliances	PureData Systems for Analytics (formerly Netezza)	Integrated, converged system built and optimized specifically for high-performance data warehousing
	IBM family of big data "accelerator" solutions	► IBM Accelerator for Social Data Analytics ► IBM Accelerator for Telco Event Data Analytics ► IBM Accelerator for Machine Data Analytics	Built to help clients accelerate their adoption of big data solutions by providing industry-specific sample solutions

[a] *Ecosystem* refers to the components and data that are involved in a Hadoop environment.

[b] Cynthia M. Saracco and Uttam Jain, "What's the big deal about Big SQL?" IBM developerWorks®, 2013. http://www.ibm.com/developerworks/library/bd-bigsql/

[c] See the description on the IBM DB2 data compression and storage optimization web page, IBM.com: http://www.ibm.com/software/data/db2/linux-unix-windows/storage-compression.html

Although these examples are taken from open source and IBM solutions, they are indicative of the ongoing evolution of technology that is helping to enable tomorrow's big data solutions.

The sections that follow describe these emerging technologies, with their functions that relate to performance and capacity.

4.2.1 Distributed processing

Distributed processing refers to computer-networking technology on multiple computers, across different locations, that are sharing computer-processing capability. This parallelism significantly reduces processing time and improves performance.

Hadoop technology, well-suited to batch-oriented, read-intensive applications, can distribute and manage data across many nodes and disks by using the following methods:

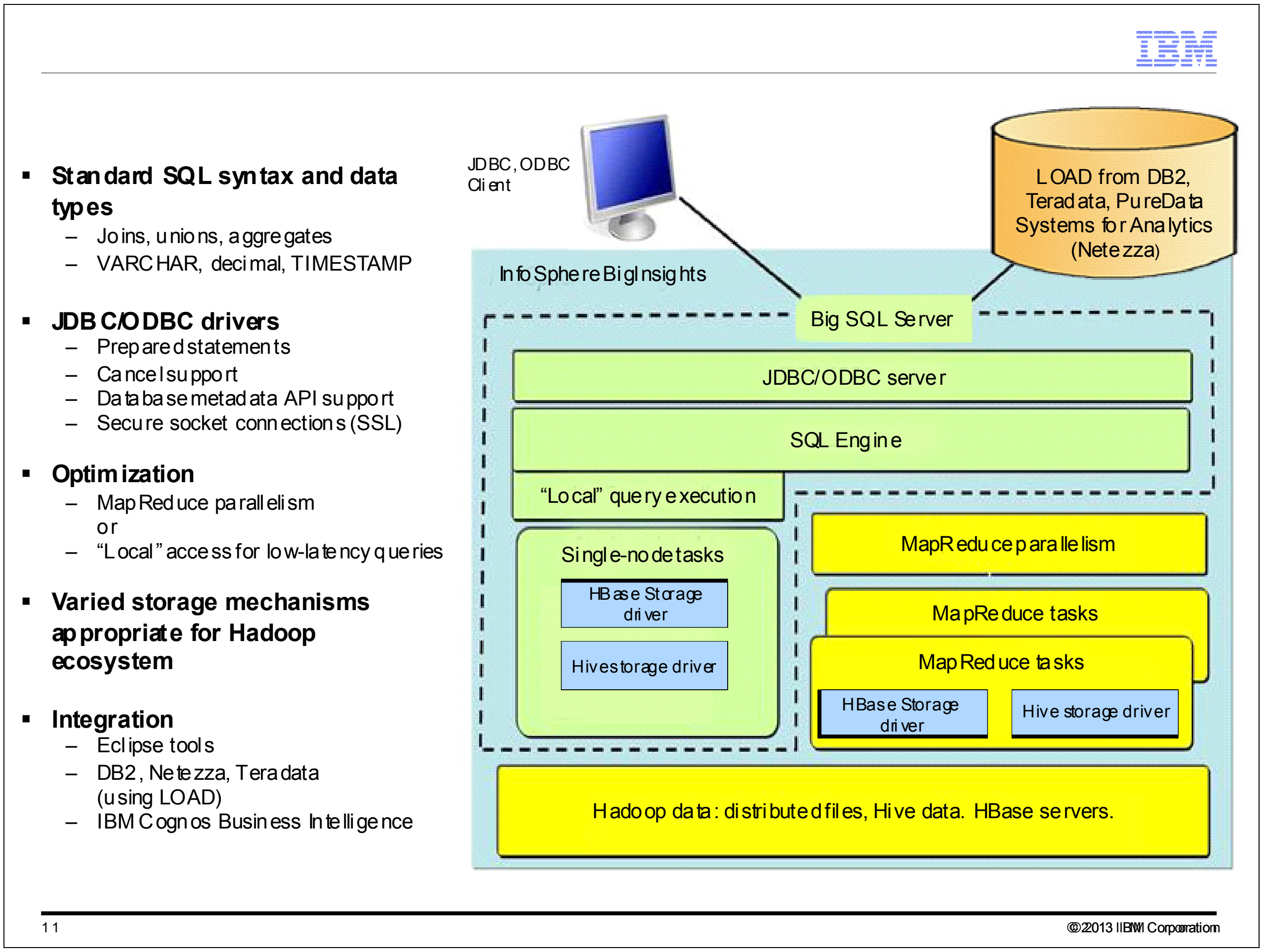
- MapReduce Engine for processing large data sets by using multiple processing nodes in both its input ("mapping") and aggregation ("reducing") tasks
- Hadoop Distributed File System (HDFS) for storing data redundantly across a cluster of processing nodes for enhanced reliability and performance

An example of big data innovation that is built on the open source tools system, including Hadoop, is IBM InfoSphere BigInsights.[2] This is a flexible, enterprise-ready analytical platform for processing large volumes of data (persistent big data, or "big data at rest") in a highly distributed and efficient manner.

The following examples include particular innovations that are provided by BigInsights:

- Big SQL is an IBM answer to the challenge of querying data from Hadoop. It offers a standard query interface to data stored in various Hadoop-based mechanisms. Figure 4-1 on page 27 shows Big SQL architecture and functions.

[2] Ayhan Onder, "IBM InfoSphere BigInsights: Smart Analytics for Big Data." PDF of presentation, 2012. ftp://ftp.software.ibm.com/software/pdf/tr/02_BigData_BigInsights.pdf



IBM

- **Standard SQL syntax and data types**
  - Joins, unions, aggregates
  - VARCHAR, decimal, TIMESTAMP
- **JDBC/ODBC drivers**
  - Prepared statements
  - Cancel support
  - Database metadata API support
  - Secure socket connections (SSL)
- **Optimization**
  - MapReduce parallelism
    or
  - "Local" access for low-latency queries
- **Varied storage mechanisms appropriate for Hadoop ecosystem**
- **Integration**
  - Eclipse tools
  - DB2, Netezza, Teradata (using LOAD)
  - IBM Cognos Business Intelligence

JDBC, ODBC Client

LOAD from DB2, Teradata, PureData Systems for Analytics (Netezza)

InfoSphere BigInsights

Big SQL Server

JDBC/ODBC server

SQL Engine

"Local" query execution

Single-node tasks

HBase Storage driver

Hive storage driver

MapReduce parallelism

MapReduce tasks

MapReduce tasks

HBase Storage driver

Hive storage driver

Hadoop data: distributed files, Hive data. HBase servers.

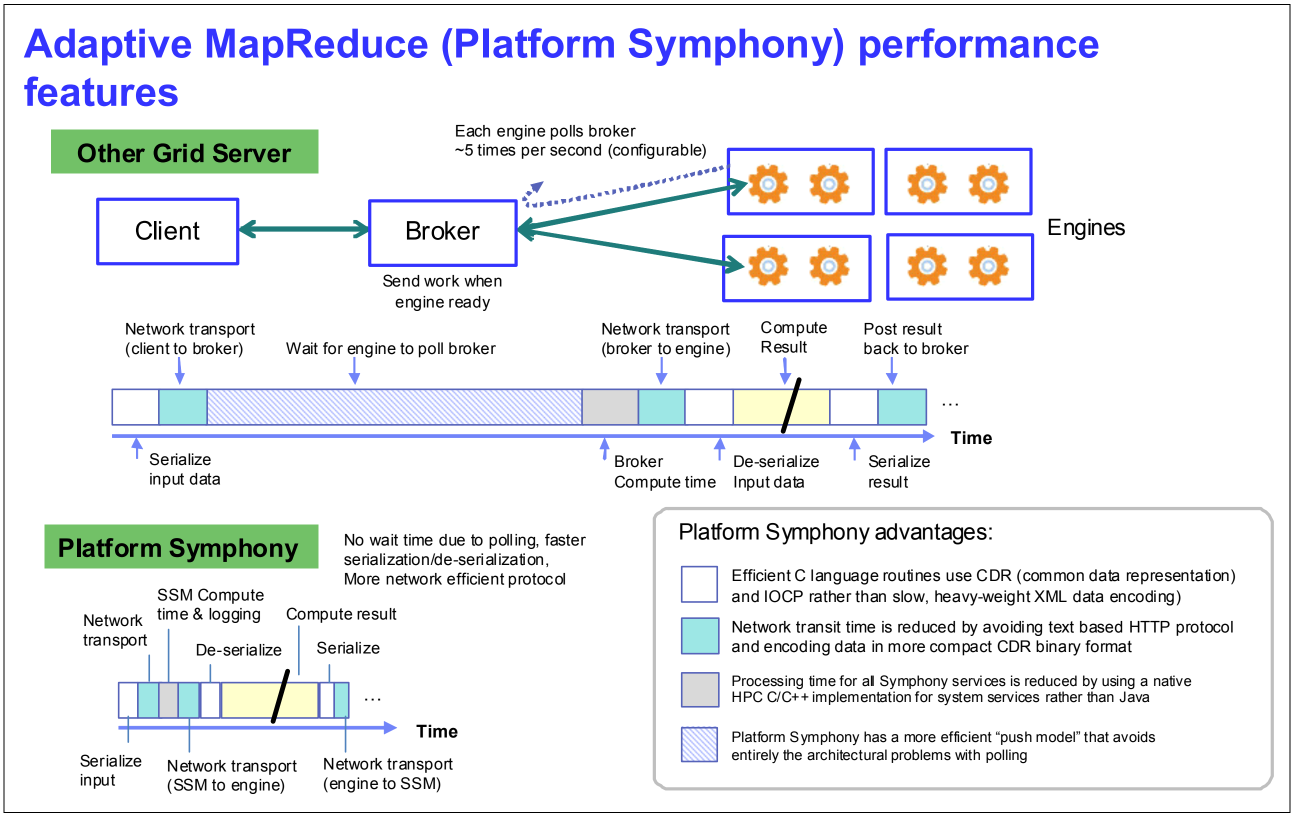
11

© 2013 IBM Corporation

Figure 4-1 Big SQL overview

Note: Low-latency refers to handling certain data elements from the time they enter the system to the time they are processed. Local access allows the processing to be quicker.

- Adaptive MapReduce in BigInsights represents an installation option for using IBM Platform Symphony® technology in place of Apache technology for MapReduce. Adaptive MapReduce can produce significant runtime performance benefits for certain workloads, particularly those that involve smaller jobs. Figure 4-2 on page 28 summarizes some of the technologies in Adaptive MapReduce that contribute to improved performance.



Adaptive MapReduce (Platform Symphony) performance features

Platform Symphony advantages:

- □ Efficient C language routines use CDR (common data representation) and IOCP rather than slow, heavy-weight XML data encoding)
- 🟦 Network transit time is reduced by avoiding text based HTTP protocol and encoding data in more compact CDR binary format
- ⬜ Processing time for all Symphony services is reduced by using a native HPC C/C++ implementation for system services rather than Java
- ▨ Platform Symphony has a more efficient "push model" that avoids entirely the architectural problems with polling

Figure 4-2 Adaptive MapReduce overview

- The IBM General Parallel File System (GPFS) File Placement Optimizer is a file system alternative to HDFS. It offers these advantages:
  - Use of local disks
  - Proven distributed file system
  - Elimination of single points of failure by using distributed metadata
  - File system access that is compliant with Portable Operating System Interface (POSIX)
  - Concurrent read/write by multiple programs
  - Support for storage pools
  - Snapshot capability
  - Security with access control list (ACL) support

4.2.2 Streams computing

Streams technology[3] enables continuous and exceptionally fast analysis of massive volumes of *information-in-motion* to help improve business insights and decision making. Big data environments must frequently meet these requirements:

- Harness and process streaming data sources
- Select valuable data and insights to be stored for further processing
- Quickly process and analyze perishable data
- Take timely action

[3] Senthil Nathan, "InfoSphere Streams, Technical Overview," IBM Research, 2009. http://ibm.co/19r63Wb

To achieve these requirements, IBM Stream Computing significantly reduces the time that is required to store valuable data and then react in real time to capture opportunities before the data expires. IBM InfoSphere Streams use streaming to address the needs of real-time big data, as shown in Figure 4-3.

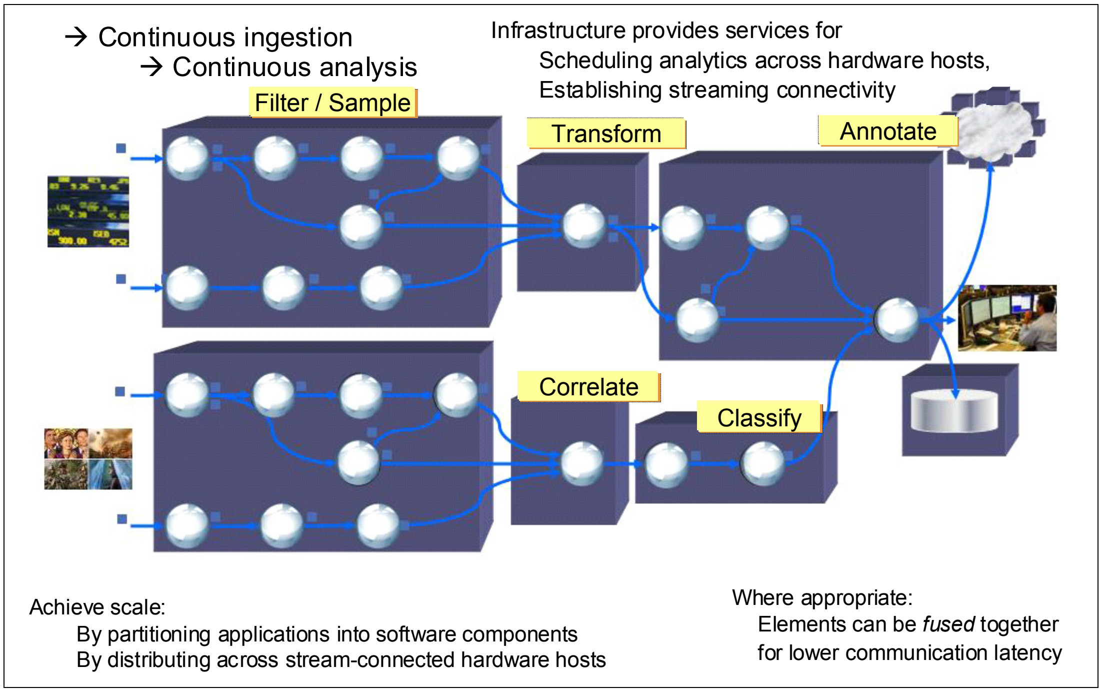

→ Continuous ingestion
→ Continuous analysis
Infrastructure provides services for
Scheduling analytics across hardware hosts,
Establishing streaming connectivity
Filter / Sample
Transform
Annotate
Correlate
Classify
Achieve scale:
By partitioning applications into software components
By distributing across stream-connected hardware hosts
Where appropriate:
Elements can be *fused* together
for lower communication latency

Figure 4-3 IBM Stream Computing components

Each software component in an IBM streams solution does accelerated processing. However, the component tasks are distributed to use parallelism, yet they are interconnected to implement the required processing sequence. With streams optimized to receive and handle multiple types of data rapidly, solutions can be designed to handle the requirements of many real-time big data environments. Otherwise, that might not be feasible. Because the data is being processed in real-time, storage requirements can often be drastically reduced.

4.2.3 Indexing

Whether referring to the keyed files and relational databases of yesterday, the search engines of today, or the big data solutions of tomorrow, *indexing* refers to the data structures built to accelerate searching, discovery, and exploration.

Many big data solutions must handle unstructured data (for example, web pages that are retrieved from a browser-based search and consist mostly of text). The indexing strategy and design is paramount for being able to achieve high performance.

Data discovery and exploration[4] is a foundation for search, discovery, and navigation systems. As implemented in IBM InfoSphere Data Explorer, the key areas that differentiate this technology are architectural in nature and inherent in the technology, not features that can be built around any indexing and search engine. They provide the following capabilities:

[4] InfoSphere Data Explorer architecture
http://public.dhe.ibm.com/common/ssi/ecm/en/imw14691usen/IMW14691USEN.PDF

- Data Explorer uses a position-based index approach that offers advantages over the traditional vector-based index approach. The position-based index contains both the keyword and positions of keywords within the original document. This makes it easier to reconstruct the contents of each document and to document subparts because position information does not have to be recovered from another source.
- Data Explorer enables index distribution (division of an index across multiple systems) to solve performance and storage problems.
- Data Explorer enables index replication (maintenance of an identical copy of indexes across multiple machines, mirroring) to use distributed processing and improve redundancy.

The result is a solution that optimizes performance, storage use, scalability, and federation of large amounts of data from multiple sources.

4.2.4 Accelerating analytics for traditional databases

Although many big data solutions use Hadoop, there are times when big data analytics and reporting need to use data that is stored in relational databases, such as an IBM DB2 database. For these situations, acceleration techniques, such as in-memory processing and data compression, can be used to more rapidly access standard relational data.

DB2 with BLU Acceleration[5] contains technology to speed analytics and reporting by using dynamic in-memory columnar methods. BLU Acceleration provides the following features:

- Actionable compression enables data to be analyzed in compressed format.
- Dynamic in-memory technology loads terabytes of data in random access memory (RAM) rather than on hard disks. This streamlines query workloads even when data sets exceed the size of the memory.
- Columnar store scans and locates the most relevant data based on columns rather than rows, which results in faster processing.
- Parallel vector processing provides multi-core and multiple data parallelism that enable you to analyze data in parallel over different processors.
- Data skipping lets you skip unnecessary processing of irrelevant or duplicate data and load only the information that needs to be analyzed.

With the use of this acceleration technology, the analysis of massive amounts of DB2 data is done efficiently and effectively without having to modify SQL or otherwise tune databases. Moreover, the compression techniques save on storage, but I/O, processor, and memory are applied in a balanced manner to speed analysis and reporting tasks.

4.2.5 Big sorting

TeraSort (a built-in Hadoop benchmark tool for sorting one terabyte of data[6]) is one of the key benchmarks for big data clients in making purchase decisions. This is because the TeraSort benchmark puts a balanced requirement on processor, memory, disk, and network, so it serves as a reliable indicator of overall system performance. In support of its analytics research mission, the IBM China Research Lab has a group that specializes in tuning Hadoop systems to optimize TeraSort results.

[5] DB2 with BLU Acceleration
http://www.ibm.com/software/data/db2/linux-unix-windows/db2-blu-acceleration/

[6] Nitin Bandugula, “Breaking the Minute Barrier for TeraSort.” *Wired* magazine, 2012.
http://www.wired.com/insights/2012/11/breaking-the-minute-barrier-for-terasort/

A full-system MapReduce optimization approach is adopted, based on correlation-based performance analysis. Because the overall performance of MapReduce applications is determined by the bottlenecks (such as the slowest map or reduce task), identifying these bottlenecks is essential for performance optimization. To identify the bottlenecks, IBM developed a performance tool to correlate different subphases (of a task), tasks (for example, map/reduce tasks), and resources (such as processor, memory, and disk) together. After bottlenecks are identified, some empirical rules can be applied to guide the practitioners to a diagnosis and help them optimize the overall performance.

By using this approach, we successfully sorted 1 TB of data in less than eight minutes on a 10-node IBM POWER7® R2 system.

For more information, see the IBM PowerLinux™ solutions web page:

http://www.ibm.com/systems/power/software/linux/powerlinux/

The performance improvement comes from different sources: About 30% from system software-level tuning (JVM, garbage collection or GC, and huge page) and about 20% from underlying hardware-level tuning (turbo mode, simultaneous multithreading, and hardware prefetching).

4.2.6 Packaged solutions

To help clients expediently deploy big data solutions, vendors frequently offer packages that incorporate relevant innovations, such as these examples.

Analytical DBMS-MPP data warehouse appliances

- IBM PureData™ Systems for Analytics[7]: Systems to deliver data warehouse services that are optimized for high-speed and peta-scale analytics, by providing built-in expertise. This includes speed to insight with built-in social data, machine data, text analytics accelerators, and speed to value with accelerated deployment. Other characteristics include no indexes or tuning, an independent data model, hardware accelerated fully parallel and optimized: Asymmetric Massively Parallel Processing (AMPP), Field Programmable Gate Array (FPGA) hardware-based acceleration, and database analytics.
- IBM PureData Systems for Hadoop: Purpose-built, standards-based, expert integrated systems that architecturally integrate BigInsights Hadoop-based software, server, and storage into a single, easy-to-manage system.

Analytic components to accelerate development and implementation

- IBM Accelerator for Social Data Analytics: A set of end-to-end applications that extract relevant information from tweets, boards, and blogs, and then build social profiles of users, based on specific use cases and industries. A typical workflow consists of importing data files and then configuring, indexing, and analyzing the data. For more information, see *Introduction to IBM Accelerators for Big Data* in the IBM InfoSphere BigInsights Information Center:

  http://ibm.co/KYcR2i

[7] IBM PureData System for Analytics N1001, powered by Netezza technology
http://public.dhe.ibm.com/common/ssi/ecm/en/imd14400usen/IMD14400USEN.PDF

- IBM Accelerator for Telecommunications Event Data Analytics: Enables telecommunications industry clients who are facing the challenge of performing real-time mediation and analytics on large volumes of Call Detail Records to import and analyze raw telecommunications data in real time, and then transform that data into meaningful and actionable insight.

  For more information, see *IBM Accelerator for Telecommunications Event Data Analytics* in the IBM InfoSphere Streams Information Center:

  `http://ibm.co/19r9V9w`

- IBM Accelerator for Machine Data Analytics: A set of end-to-end applications that help import, extract, index, transform, and analyze data to perform faceted search and event and pattern correlation and then make informed decisions that are based on the intelligence that is contained in log and data files. A typical introductory workflow consists of organizing and importing batches of data, and then extracting, indexing, searching, transforming, and analyzing the data. For more information, see *Introduction to IBM Accelerators for Big Data* in the IBM InfoSphere BigInsights Information Center:

  `http://ibm.co/KYcR2i`

5

Summary

When big data is considered from the perspective of performance and capacity, the predominant technical requirement of IT capabilities is massive scalability. Although continuing improvements in the cost, size, speed, and efficiency of processing, memory, and storage hardware have helped to some extent, it is mostly innovations in using distributed processing that have provided the most dramatic increases in big data capabilities. In addition to refinements in how data is stored, indexed, accessed, manipulated, analyzed, and distributed by using the available hardware resources, “scaling out” has gone a long way toward making big data solutions both feasible and manageable from a performance and capacity perspective.

In addition to technical innovation, the story of big data is also one of cooperation. For example, the open source community’s contribution of Apache Hadoop and its related tools provides a strong base set of distributed big data capabilities for the rest of industry to use, enhance, and augment. Likewise, IBM and other vendors enhance their offerings and create new ones. These products accelerate data access, exploration, and analytics by using both existing data platforms and emerging platforms, such as the Hadoop ecosystem. They also facilitate customer adoption of big data approaches to address pressing business challenges.

This paper covered the implications for performance and capacity of the popular “four Vs” of big data: volume, velocity, variety, and veracity. More recently, two more Vs have been suggested by various commentators:

- *Variability* of data injection into big data systems
- *Value* of big data

Variability means that big data systems need to display the same kind of elasticity that is required of cloud computing and other virtualized environments. Therefore, virtualization continues to play a key role.

The value of big data is realized by those enterprises that embrace data as a competitive differentiator and respond accordingly. Given this value proposition, it is the availability of high-performing, cost-effective, and scalable big data solutions that makes the business case for big data viable.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this paper.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this paper. Note that some publications referenced in this list might be available in softcopy only.

- *Addressing Data Volume, Velocity, and Variety with IBM InfoSphere Streams V3.0*, SG24-8108
- *Implementing IBM InfoSphere BigInsights on IBM System x*, SG24-8077
- *Smarter Analytics: Information Architecture for a New Era of Computing*, SG24-5012

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Online resources

These websites are also relevant as further information sources:

- IBM big data platform web page

  http://www.ibm.com/software/data/bigdata/
- Demystifying Big Data: A Practical Guide to Transforming the Business of Government. TechAmerica Foundation, not dated.

  http://public.dhe.ibm.com/common/ssi/ecm/en/iml14336usen/IML14336USEN.PDF
- T. H. Davenport, P. Barth, and R. Bean, “How ‘Big Data’ Is Different.” MIT Sloan Management Review, Fall 2012, Vol. 54, No. 1.

  http://www.stevens.edu/howe/sites/default/files/MIT-SMR%20How%20Big%20Data%20is%20Different.pdf
- Jeff Jonas on analytics, IBM Innovation explanations.

  http://www.ibm.com/smarterplanet/us/en/innovation_explanations/article/jeff_jonas.html
- DB2 with BLU Acceleration

  http://www.ibm.com/software/data/db2/linux-unix-windows/db2-blu-acceleration/

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Performance and Capacity Implications for Big Data

Redpaper™

Overview and benefits of big data

Critical factors in managing big data successfully

Emerging technologies that address big data challenges

Big data solutions enable us to change how we do business by exploiting previously unused sources of information in ways that were not possible just a few years ago. In IBM Smarter Planet terms, big data helps us to change the way that the world works.

The purpose of this IBM Redpaper publication is to consider the performance and capacity implications of big data solutions, which must be taken into account for them to be viable. This paper describes the benefits that big data approaches can provide. We then cover performance and capacity considerations for creating big data solutions. We conclude with what this means for big data solutions, both now and in the future.

Intended readers for this paper include decision-makers, consultants, and IT architects.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks