Lenovo Big Data Validated Design for IBM Db2 Warehouse on ThinkSystem Servers

Last update: **16 March 2018**
Version 1.0

Describes the reference architecture for high performance infrastructure for IBM Db2 Warehouse

Solution based on the powerful, versatile Lenovo ThinkSystem SR650 servers powered by Intel® Xeon® Scalable Processors

Deployment considerations for high-performance, cost-effective and scalable solutions

Uses Intel NVMe storage and Lenovo network devices to deliver very high performance

Lenovo

IBM

Intel

LENOVO PRESS

Table of Contents

1 Introduction

This document describes the reference architecture for the Lenovo Big Data Validated Design for IBM Db2 Warehouse with ThinkSystem Servers. It provides a predefined and optimized hardware infrastructure for high performance implementation of IBM Db2 Warehouse software. This reference architecture provides planning, design considerations, and best practices for implementing IBM Db2 Warehouse with Lenovo and Intel products.

The Lenovo, IBM and Intel teams worked together on this document and the reference architecture described herein was developed and validated in a joint engineering project.

With ever increasing amounts of data being made available to an enterprise, the challenge of deriving the most value from it has become very important. This task requires the use of suitable analytics software running on a tuned hardware platform. Running Db2 Warehouse for Data Marts and Enterprise Data Warehouse (EDW) use cases as well as incorporating it in Data Lakes and driving EDW modernization are areas of significant interest and focus. The IBM Db2 Warehouse solution described in this document is very well suited for implementing the infrastructure to support these modern analytics initiatives.

IBM Db2® Warehouse is an analytics data warehouse that you deploy by using a Docker container, allowing you control over data and applications, but simplicity in terms of deployment and management. Db2 Warehouse offers in-memory BLU processing technology and in-database analytics, plus scalability and performance through both the single node (SMP) and multi-node (MPP) architecture.

An analytics system must be balanced to deliver results that enterprises demand today to meet their needs for information and insights. Achieving extremely high performance requires that high performance processors, large memory capacity and low-latency, high-bandwidth storage and networking are employed. The Lenovo servers used in this solution are powered by Intel Xeon Scalable Processor family processors and Intel NVMe solid state drives (SSDs). Furthermore, as enterprises exploit the value of analytics and deploy hardware platforms with high performance CPU’s and NVMe storage, the system connectivity requirements must also be addressed.

The predefined configuration provides a baseline configuration for Db2 Warehouse solution which can be modified based on specific customer requirements such as lower cost, different storage needs and increased reliability.

The intended audience of this document is IT professionals, technical architects, sales engineers, and consultants to assist in planning, designing, and implementing the big data solution with Lenovo hardware. It is assumed that you are familiar with data warehouse, Spark and SQL concepts. For more information, see “Resources” on page 19.

2 Business problem and business value

This section describes business challenges faced by big data environments and the value provided by the IBM Db2 Warehouse solution used to address the business challenges.

2.1 Business problem

The world is well on its way to generate more than 40 million TB of data by 2020. In all, 90% of the data in the world today was created in the last two years alone. This data comes from everywhere, including sensors that are used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone global positioning system (GPS) signals. This data is big data.

Harnessing value from all this data is a key challenge for all enterprises. The need for extracting valuable insights from the data in a timely manner and under stringent performance constraints is a major problem. The modernization of the workload applications as well as the underlying IT infrastructure is driving additional requirement for fast access to enterprise data. Installation and management of data repositories like data warehouses and data lakes is both time and effort intensive. Transferring large amounts of data to an analytics engine every time a query needs to be run is very expensive.

2.2 Business value

Big data is more than a challenge; it is an opportunity to derive insight from new and emerging types of data to make your business more intelligent. Big data also is an opportunity to answer questions that, in the past, were beyond reach. The IBM Db2 Warehouse solution described in this document addresses the key challenges outlined above by delivering value based on the following capabilities:

- Deploy a pre-configured data warehouse in minutes on Docker container supported infrastructure of choice, with elastic scaling and ease of updates/upgrades to continually meet service levels.
- Gain insight when applying in-database analytics where the data resides.
- Use Spark and IBM BLU Acceleration in-memory SQL columnar processing with a MPP cluster architecture to speed up complex queries and predictive model building, testing, and deployment.
- With just a few clicks, unstructured data sources are automatically transformed into a structured format for analysis – Twitter data, Open Data, geospatial data, and more.
- A common SQL analytics engine across public and private cloud, compatible with on-premises data warehouses, enables workloads to be moved to the right location with ease.

This reference architecture document describes the performance and scalability benefits of deploying IBM Db2 Warehouse on the Lenovo servers using Intel NVMe drives and Mellanox network interface cards.

3 Requirements

The functional and non-functional requirements for this reference architecture are desribed in this section.

3.1 Functional requirements

A big data solution supports the following key functional requirements:

- Ability to handle various workloads, including batch and real-time analytics
- Industry-standard interfaces, such as SQL so that applications can reach the data easily
- Ability to handle large volumes of data of various data types
- Various client interfaces

3.2 Non-functional requirements

Customers require their big data solution to be easy, dependable, and fast. The following non-functional requirements are key:

- Easy:
  - Ease of development
  - Easy management at scale
  - Advanced job management
  - Multi-tenancy
  - Easy to access data by various user types
- Dependable:
  - Data protection with snapshot and mirroring
  - Automated self-healing
  - Insight into software/hardware health and issues
  - High availability (HA) and business continuity
- Fast:
  - Superior performance
  - Scalability
- Secure and governed:
  - Strong authentication and authorization
  - Kerberos support
  - Data confidentiality and integrity

4 Architectural overview

The IBM Db2 Warehouse solution is based on a flexible and scalable reference architecture. The primary hardware building block is the data node implemented on ThinkSystem SR650 servers. A cluster of SR650 servers are connected together to meet the desired total memory size required to deliver the best performance for IBM Db2 Warehouse solution. A clustered file system is used to share data across all the servers in the cluster.

Figure 1 shows the architecture overview of the IBM Db2 Warehouse reference architecture that uses Lenovo ThinkSystem hardware infrastructure.

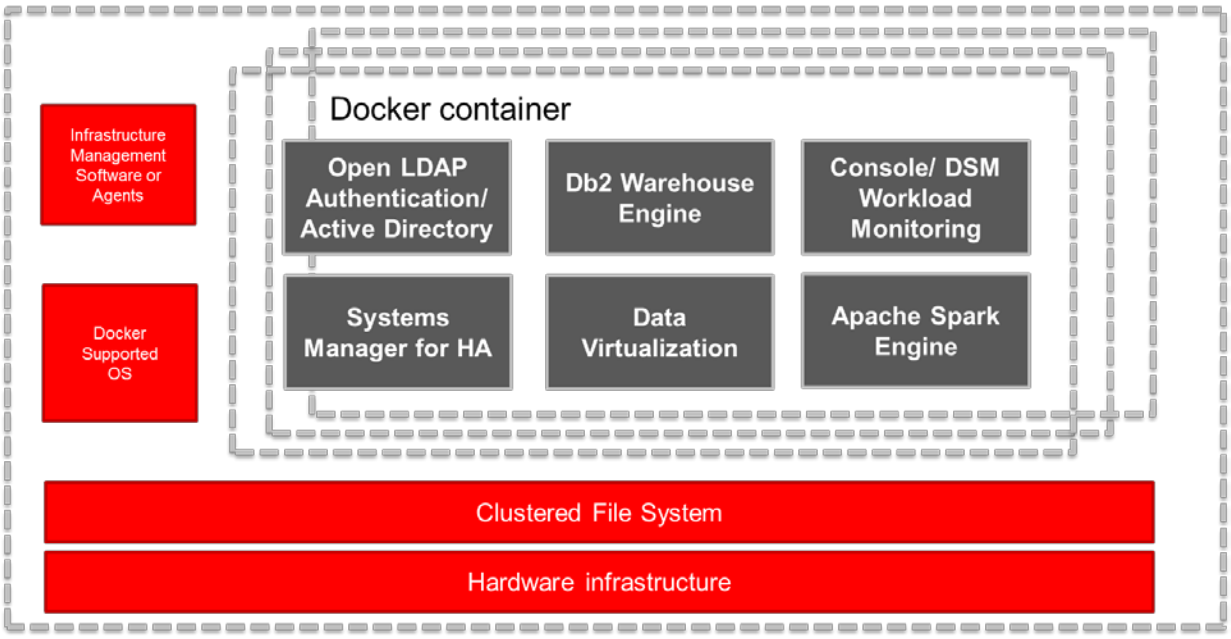

Figure 1. IBM Db2 Warehouse reference architecture overview.

5 Component model

This section describes the high-level component model of the IBM Db2 Warehouse solution as shown in Figure 2.

5.1 IBM Db2 Warehouse overview

IBM Db2® Warehouse is an analytics data warehouse that you deploy by using a Docker container, allowing you control over data and applications, but simplicity in terms of deployment and management. Db2 Warehouse offers in-memory BLU processing technology and in-database analytics, plus scalability and performance through the MPP architecture.

Db2 Warehouse might be a suitable option if any of the following criteria apply to you:

- Your data must stay on premises because of privacy requirements.
- You want the flexibility of the cloud without giving up control over your data.
- You have workloads to move between a public cloud or appliance and a private cloud.
- You are considering a hybrid architecture to modernize your data warehouse.
- You want to use a private cloud as a first step to public cloud deployment.

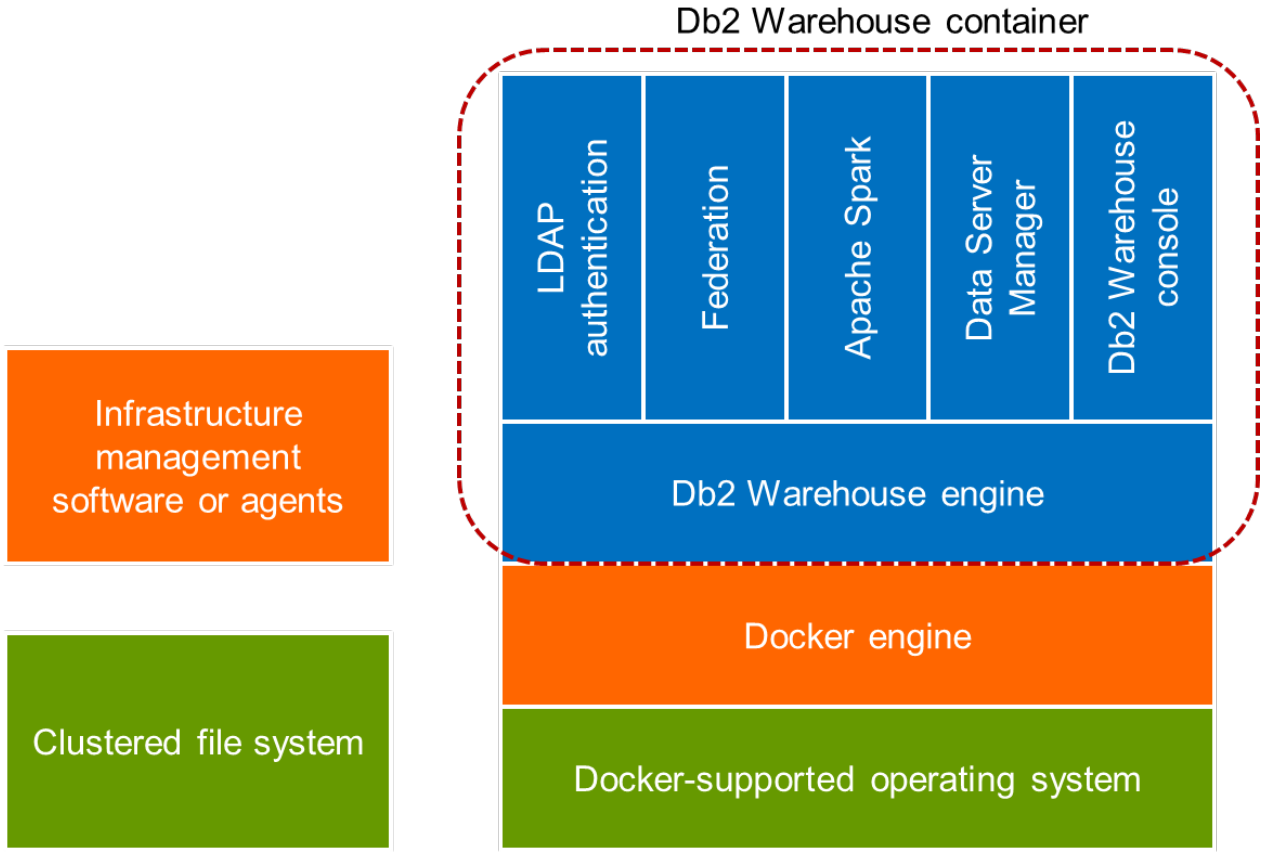

Figure 2. IBM Db2 Warehouse component model overview.

The IBM Db2 Warehouse container includes the following components:

- Db2 Warehouse engine
- LDAP authentication
- Federation
- Apache Spark engine
- Data Server Manager
- Db2 Warehouse console.

Apache Spark engine provides a framework for in-memory processing and also allows use of Spark extensions such as those for SQL and Machine Learning use cases.

The IBM Db2 Warehouse solution employs a clustered file system for sharing data across the nodes. In this reference architecture, IBM Spectrum Scale is used as the clustered file system.

You can deploy Db2 Warehouse in a wide range of environments, from a basic laptop for development purposes, all the way to a large production cluster. You can choose either a single-node (SMP) deployment or a multi-node (MPP) deployment. (On Windows and Macintosh, only SMP deployments are supported.) An MPP deployment has a minimum of three nodes and a maximum of either 24 or 60 nodes. The maximum depends on the number of data partitions that were allocated when you deployed.

The containerization technology that Db2 Warehouse uses makes deployment fast (typically fewer than 30 minutes for an MPP cluster and significantly less for SMP) and simple (usually only one or two commands are required to download and initialize the image). As you can see in Figure 2, the Db2 Warehouse container is lightweight because it doesn't contain a guest operating system or a hypervisor, as with a VM. The Db2 Warehouse software stack is isolated in its own container but allows you to use your existing infrastructure and cloud management or monitoring tools.

5.2 IBM Spectrum Scale overview

IBM Spectrum Scale™ is a cluster file system that provides concurrent access to a single file system or set of file systems from multiple nodes. The nodes can be SAN attached, network attached, a mixture of SAN attached and network attached, or in a shared nothing cluster configuration. This enables high performance access to this common set of data to support a scale-out solution or to provide a high availability platform.

IBM Spectrum Scale has many features beyond common data access including data replication, policy based storage management, and multi-site operations. You can create a cluster of AIX® nodes, Linux nodes, Windows server nodes, or a mix of all three. IBM Spectrum Scale can run on virtualized instances providing common data access in environments, leverage logical partitioning, or other hypervisors. Multiple IBM Spectrum Scale clusters can share data within a location or across wide area network (WAN) connections.

IBM Spectrum Scale, the follow-on to IBM GPFS, is a high-performance solution for managing data at scale with the distinctive ability to perform archive and analytics in place.

IBM Spectrum Scale has the following features:

- Uses Declustered RAID, where data and parity information as well as Spare Capacity is distributed across all disks
- Rebuilds with Declustered RAID are faster:
    - Traditional RAID would have one LUN fully busy resulting in slow rebuild and high impact overall
    - Declustered RAID rebuild activity spreads the load across many disks resulting in faster rebuild and less disruption to user programs
    - Declustered RAID minimizes critical data exposed to data loss in case of a second failure.
- 2-fault / 3-fault tolerance and mirroring: 2- or 3-fault-tolerant Reed-Solomon parity encoding as well as 3- or 4-way mirroring provides data integrity, reliability and flexibility
- End-to-end checksum:
    - Helps detect and correct off-track I/O and dropped writes
    - Disk surface to GPFS user/client provides information to help detect and correct write or I/O errors
- Disk hospital - asynchronous, global error diagnosis:
    - If there is a media error, information provided helps in verifying and restoring a media error.
    - If there is a path problem, information can be used to attempt alternate paths.
    - Disk tracking information helps track disk service times, which is useful in finding slow disks so they can be replaced.
    - Multipathing: Performed automatically by Spectrum Scale, so no multipath driver is needed.
    - Supports a variety of file I/O protocols:
        - POSIX, GPFS, NFS v4.0, SMB v3.0
        - Big data and analytics: Hadoop MapReduce
        - Cloud: OpenStack Cinder (block), OpenStack Swift (object), S3 (object)
- Supports cloud object storage:
    - IBM Cloud Storage System (Cleversafe)
    - Amazon S3
    - IBM SoftLayer Native Object
    - OpenStack Swift
    - Amazon S3 compatible providers

6 Operational model

This section describes the hardware infrastructure aspects of the IBM Db2 Warehouse reference architecture. To support different customer environments, different configurations are provided for supporting different amounts of data sizes and performance levels.

6.1 Hardware description

This reference architecture uses Lenovo servers SR650 (2U) servers and Lenovo RackSwitch G8052 and NE10032 top of rack switches.

6.1.1 Lenovo ThinkSystem SR650 Server

The Lenovo ThinkSystem SR650 is an ideal 2-socket 2U rack server for small businesses up to large enterprises that need industry-leading reliability, management, and security, as well as maximizing performance and flexibility for future growth. The SR650 server is particularly suited for big data applications due to its rich internal data storage, large internal memory and selection if high performance Intel processors. It is also designed to handle general workloads, such as databases, virtualization and cloud computing, virtual desktop infrastructure (VDI), enterprise applications, collaboration/email, and business analytics.

The SR650 server supports:

- Up to two Intel® Xeon® Scalable Processors
- Up to 1.5 TB 2666 MHz TruDDR4 memory (support for up to 3 TB is planned for future),
- Up to 24x 2.5-inch or 14x 3.5-inch drive bays with an extensive choice of NVMe PCIe SSDs, SAS/SATA SSDs, and SAS/SATA HDDs
- Flexible I/O Network expansion options with the LOM slot, the dedicated storage controller slot, and up to 6x PCIe slots

Figure 3. Lenovo ThinkSystem SR650

Combined with the Intel® Xeon® Scalable Processors (Bronze, Silver, Gold, and Platinum), the Lenovo SR650 server offers an even higher density of workloads and performance that lowers the total cost of ownership (TCO). Its pay-as-you-grow flexible design and great expansion capabilities solidify dependability for any kind of workload with minimal downtime.

The SR650 server provides high internal storage density in a 2U form factor with its impressive array of workload-optimized storage configurations. It also offers easy management and saves floor space and power consumption for most demanding use cases by consolidating storage and server into one system.

This reference architecture recommends the storage-rich ThinkSystem SR650 for the following reasons:

- **Storage capacity**: The nodes are storage-rich. Each of the 14 configured 3.5-inch drives has raw capacity up to 10 TB and each, providing for 140 TB of raw storage per node and over 2000 TB per rack.
- **Performance**: This hardware supports the latest Intel® Xeon® Scalable processors and TruDDR4 Memory.
- **Flexibility**: Server hardware uses embedded storage, which results in simple scalability (by adding nodes).
- **PCIe slots**: Up to 7 PCIe slots are available if rear disks are not used, and up to 3 PCIe slots if the Rear HDD kit is used. They can be used for network adapter redundancy and increased network throughput.
- **Higher power efficiency**: Titanium and Platinum redundant power supplies that can deliver 96% (Titanium) or 94% (Platinum) efficiency at 50% load.
- **Reliability:** Outstanding reliability, availability, and serviceability (RAS) improve the business environment and helps save operational costs

For more information, see the Lenovo ThinkSystem SR650 Product Guide:

https://lenovopress.com/lp0644-lenovo-thinksystem-sr650-server

6.1.2 Lenovo RackSwitch G8052

The Lenovo networking RackSwitch G8052 (as shown in Figure 4) is an Ethernet switch that is designed for the data center and provides a simple network solution. The Lenovo RackSwitch G8052 offers up to 48x 1 GbE ports and up to 4x 10 GbE ports in a 1U footprint. The G8052 switch is always available for business-critical traffic by using redundant power supplies, fans, and numerous high-availability features.

Figure 4. Lenovo RackSwitch G8052

Lenovo RackSwitch G8052 has the following characteristics:

- A total of **48x 1 GbE** RJ45 ports
- **Four 10 GbE** SFP+ ports
- Low 130W power rating and variable speed fans to reduce power consumption

For more information, see the Lenovo RackSwitch G8052 Product Guide:
https://lenovopress.com/tips1270-lenovo-rackswitch-g8052

6.1.3 Lenovo RackSwitch NE10032 - Cross-Rack Switch

The Lenovo ThinkSystem NE10032 RackSwitch that uses 100 Gb QSFP28 and 40 Gb QSFP+ Ethernet technology is specifically designed for the data center. It is ideal for today's big data workload solutions and is an enterprise class Layer 2 and Layer 3 full featured switch that delivers line-rate, high-bandwidth switching, filtering and traffic queuing without delaying data. Large data center-grade buffers help keep traffic moving, while the hot-swap redundant power supplies and fans (along with numerous high-availability features) help provide high availability for business sensitive traffic.

The NE10032 RackSwitch has 32x QSFP+/QSFP28 ports that support 40 GbE and 100 GbE optical transceivers, active optical cables (AOCs), and direct attach copper (DAC) cables. It is an ideal cross-rack aggregation switch for use in a multi rack big data cluster.

Figure 5: Lenovo ThinkSystem NE10032 cross-rack switch

For further information on the NE10032 switch, visit this link:

https://lenovopress.com/lp0609-lenovo-thinksystem-ne10032-rackswitch

6.2 Cluster nodes

The IBM Db2 Warehouse reference architecture is implemented on a set of nodes that make up a cluster of data nodes and optional edge nodes. Data nodes use ThinkSystem SR650 servers. Data nodes run data services for storing and processing data. Edge node (optional) acts as a boundary between the cluster and the outside (client) environment.

6.2.1 Data nodes

Table 1 lists the recommended system components for data nodes in this reference architecture.

Table 1. Data node configuration

Component	Worker node configuration
Server	ThinkSystem SR650
Processor	2x Intel® Xeon® processors: 6140 Gold, 18-core 2.3Ghz 6152 Gold, 22-core, 2.1GHz 8170 Platinum, 26-core, 2.1GHz
Memory	384 GB to 1.5 TB: 24 x 64GB 2400MHz RDIMM

Disk (OS)	Dual M.2 128GB SSD with RAID1
Flash storage (data)	4x 3.2 TB Intel® SSD DC P4600 Series SSDs (AIC) 4x 3.2 TB Intel® SSD DC P4600 Series SSDs (2.5”)
HDD controller	OS: M.2 RAID1 mirror enablement kit HDFS: ThinkSystem 430-16i 12Gb HBA
Hardware management network adapter	Integrated XCC management controller - dedicated 1Gb or shared LAN port
Data network adapter	100GbE adapter (Mellanox ConnectX-4 EN)

The Intel® Xeon® Scalable Processor family processors recommended in Table 1 will provide a balance in performance vs. cost for data nodes. Processors with different core count and frequency are available for matching the compute intensity required by various workloads. The memory capacity can also be adjusted based on cost and performance considerations. Each worker node in the reference architecture has internal directly attached storage. External storage is not used in this reference architecture.

Mellanox ConnectX-4 Ethernet Adapters deliver highest performance network for Big Data workloads with 100Gb/s server connectivity:

- Highest performing network for applications requiring high bandwidth, low latency and high message rate
- World-class cluster, network, and storage performance
- Advanced hardware offloads for IP packet processing
- Hardware-based security and isolation for Spark workloads in containers
- End-to-end QoS and congestion control
- Efficient I/O consolidation, lowering data center costs and complexity

Intel® Solid State Drive DC P4600 Series AIC (Add-in Card) and provide the highest2.5” form factors both include:

- Consistently high IOPS and throughput
- Sustained low latency
- Variable Sector Size and End-to-End data path protection
- Power loss protection capacitor self-test
- Out of band management
- Thermal throttling and monitoring

Advantages of using Intel® NVMe™ storage – The Intel® Solid State Drive (SSD) Data Center Family for PCIe® brings extreme data throughput directly to Intel® Xeon® processors with up to six times faster data transfer speed than 6 Gbps SAS/SATA SSDs. The performance and most flexible solution for high-of a single drive from the Intel SSD Data Center Family for PCIe®, specifically the Intel® SSD DC P4600 Series (450K

IOPS), can replace the performance, Web 2.0, Cloud, data analytics, database, and storage platforms.

ConnectX-4 adapters provide robust high-speed access to the 7 SATA SSDs aggregated through an unmatched combination of 100Gb/s bandwidth in a single port, with the lowest available latency, 150 million messages per second and application hardware offloads, addressing both today'sHBA (~500K IOPS). The P4600 Series is a PCIe® Gen3 SSD architected with the new high performance controller interface – NVMe™ (Non-Volatile Memory Express™) delivering leading performance, low latency and the next generation's compute and storage data center demands Quality of Service.

The number of data nodes that are required within an IBM Db2 Warehouse cluster depends on the client requirements. Such requirements might include the size of a cluster, the size of the user data, the data compression ratio, workload characteristics, and data ingest.

A minimum of three data nodes are required for MPP configurations. Three data nodes should be used for test or POC environments only.

6.2.2 Edge nodes (optional)

The optional edge node acts as a boundary between the IBM Db2 Warehouse cluster and the outside (client) environment. The edge node is used for data ingest, which refers to routing data into the cluster through the data network of the reference architecture. Edge nodes can be Lenovo ThinkSystem SR650 servers, other Lenovo servers, or other client-provided servers.

6.3 Systems management

Systems management of a cluster includes Operating System and hardware management.

Hardware management uses the Lenovo XClarity™ Administrator, which is a centralized resource management solution that reduces complexity, speeds up response and enhances the availability of Lenovo server systems and solutions. XClarity™ is used to install the OS onto new worker nodes; update firmware across the cluster nodes, record hardware alerts and report when repair actions are needed.

Figure 10 shows the Lenovo XClarity™ Administrator interface in which servers, storage, switches and other rack components are managed and status is shown on the dashboard. Lenovo XClarity™ Administrator is a virtual appliance that is quickly imported into a server-virtualized environment.

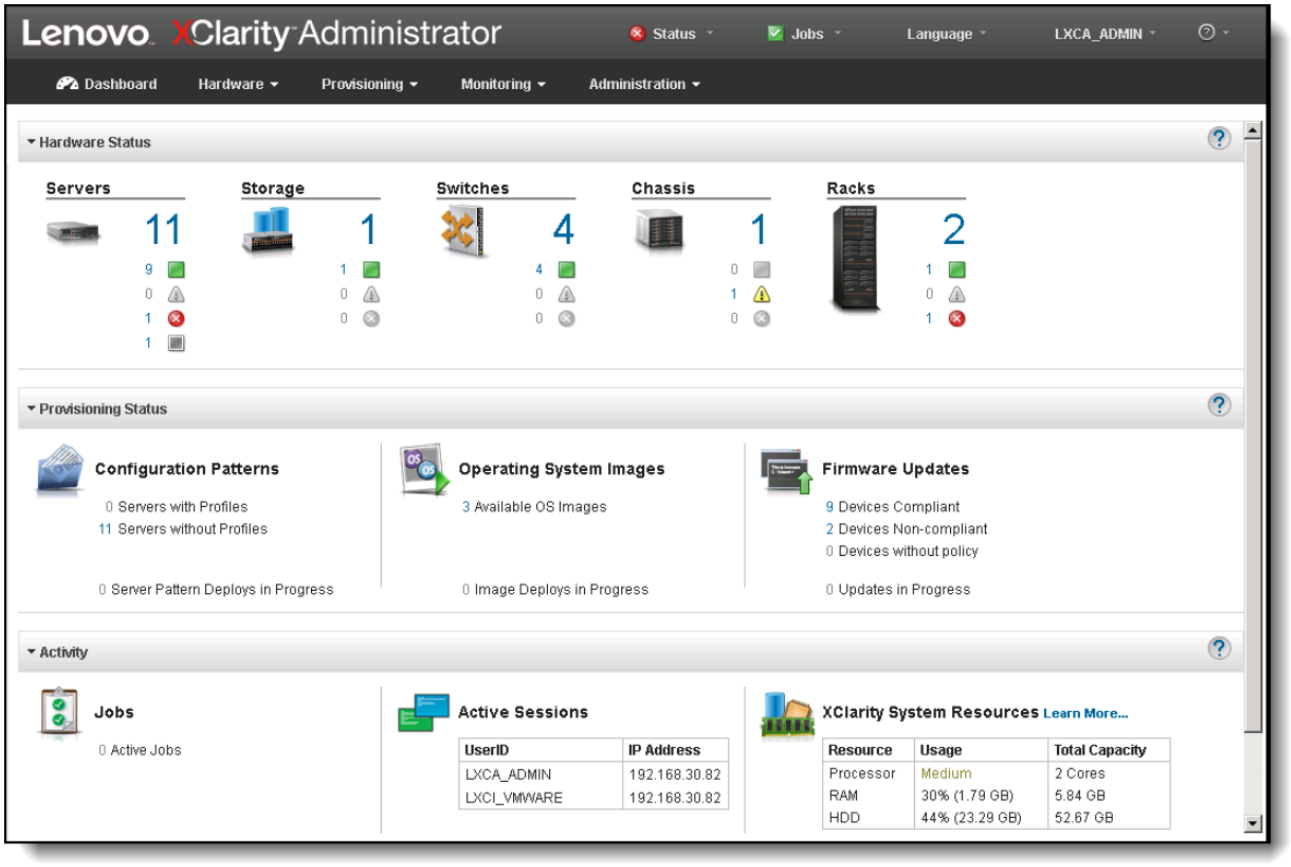

Figure 6: XClarity™ Administrator interface

In addition, xCAT provides a scalable distributed computing management and provisioning tool that provides a unified interface for hardware control, discovery and operating system deployment. It can be used to facilitate or automate the management of cluster nodes. For more information about xCAT, see “Resources” section.

6.4 Networking

Regarding networking, the reference architecture specifies two networks: a data network and an administrative or management network.

6.4.1 Data network

The current design is a single switch connected to a single port of a dual PCI Gen3.0 x16 100GbE card. This is important to remember that the maximum bandwidth of a single card is ~ 112 Gb. This means that active/active on a single card will exhaust the resources of the PCI bus before reaching 2x 100GbE. As the test detailed in this report was carried out with a single switch, a basic level of high availability can be achieved with Active/Standby and a second switch. There would need to be interconnect between the two switches so that if a port or cable failed the new active port would not be isolated from the other nodes. A complete failure of the switch would result that all the standby ports would become active and traffic would follow.

6.4.2 Hardware management network

The hardware management network is a 1 GbE network that is used for in-band operating system administration and out-of-band hardware management. In-band administrative services, such as SSH or Virtual Network Computing (VNC) that is running on the host operating system enables cluster nodes to be administered. Using the integrated management modules II (IMM2) within the ThinkSystem SR650 server, out-of-band management enables the hardware-level management of cluster nodes, such as node deployment or basic input/output system (BIOS) configuration.

Based on customer requirements, the administration links and management links can be segregated onto separate VLANs or subnets. The administrative or management network is typically connected directly to the customer's administrative network. When the in-band administrative services on the host operating system are used, the cluster is configured to use the data network only.

The reference architecture requires one 1 Gb Ethernet top-of-rack switch for the hardware management network. This rack switch for the hardware management network is connected to each of the nodes in the cluster by using two physical links (one for in-band operating system administration and one link for out-of-band IMM2 hardware management). On the nodes, the administration link connects to port 1 on the integrated 1 GBaseT adapter and the management link connects to the dedicated IMM2 port.

7 Deployment considerations

This section describes various other considerations for deploying the IBM Db2 Warehouse solution.

The predefined configurations represent a set of baseline configurations that can be implemented as specified or modified based on specific client requirements, such as lower cost, improved performance, and increased reliability.

When you consider modifying the predefined configuration, you must understand key aspects of how the cluster will be used. In terms of data, you must understand the current and future total data to be managed, the size of a typical data set, and whether access to the data will be uniform or skewed. In terms of ingest, you must understand the volume of data to be ingested and ingest patterns, such as regular cycles over specific time periods and bursts in ingest. Consider also the data access and processing characteristics of common jobs and whether query-like frameworks are used.

When designing the IBM Db2 Warehouse cluster infrastructure, we recommend conducting the necessary testing and proof of concepts against representative data and workloads to ensure that the proposed design will achieve the necessary success criteria. The following sections provide information about customizing the predefined configuration. When considering customizations to the predefined configuration, work with a systems architect who is experienced in designing the IBM Db2 Warehouse cluster infrastructures.

7.1 Designing for lower cost

There are several key modifications that can be made to lower the cost of a IBM Db2 Warehouse reference architecture solution. When lower-cost options are considered, it is important to ensure that customers understand the potential lower performance implications of a lower-cost design. A lower-cost version of the IBMDb2 Warehouse reference architecture can be achieved by using lower-cost node processors, reducing the amount of memory capacity per data node and using lower-cost storage drives.

The node processors can be substituted with other processors in the Intel Xeon SP family. Selecting a different processor may lead to a lower frequency memory, which can also lower the per-node cost of the solution.

The use of a smaller memory capacity per data node can provide a lower-cost design. However, the performance of a cluster with data nodes using the lower memory capacity can be significantly lower. Testing during proof-of-concept evaluation should be done with real user data to understand the performance implications.

The storage drives on a data node can be changed to a lower-capacity NVMe drive or a standard SSD. The impact on the performance of data nodes using these lower-cost storage options should be evaluated during proof-of-concept testing as mentioned before.

7.2 Estimating cluster size

When estimating storage space within an IBM Db2 Warehouse cluster, the following considerations are useful:

- For improved performance, most of the working set of active data should reside in memory.

- Db2 Warehouse is built with a columnar processing engine. Active working data set estimation includes active columns and rows.
- Compression ratio is an important consideration in estimating disk space and can vary greatly based on file contents.
- Working memory size is estimated based on desired concurrency level and mix of query types.

Based on these considerations, the total memory space and the required number of nodes can be estimated by using the following equations:

*Buffer pool size = (User data, uncompressed) * (% active columns * % active rows) / (compression ratio)*

Total memory size = Buffer pool size + Working memory size

Total required data nodes = (Total memory space) / (Memory per node)

The total number of data nodes should be selected based on supporting recommended splitting of the active data set. Finally, you should also consider future growth requirements when estimating memory space.

As an example, consider a user with uncompressed data size of 40 TB. Assume that 30% of the rows in the data set are active and queries generally access no more than 50% of the columns. This means that active uncompressed data size is 40 * 0.3 * 0.5 = 6 TB. Further assume that a quick check of the data shows compression rate of 6x, making compressed data size as 1 TB. Finally, assuming an equal working memory size, the today required memory size is 2 TB. Using data nodes with 768 GB of memory, three nodes will be required to provide sufficient capacity for in-memory processing.

7.3 Scaling considerations

When the capacity of the existing infrastructure is reached, the cluster can be scaled out by adding more data nodes and, if necessary, management nodes. As the capacity of existing racks is reached, new racks can be added to the cluster. Some workloads might not scale linearly.

When you design a new IBM Db2 Warehouse solution reference architecture implementation, future scale out is a key consideration in the initial design. You must consider the two related aspects of networking and management. Both of these aspects are critical to cluster operation and become more complex as the cluster infrastructure grows.

The networking model that is described in the section “Networking” on page 13 is designed to provide robust network interconnection of racks within the cluster. As more racks are added, the predefined networking topology remains balanced and symmetrical. If there are plans to scale the cluster beyond one rack, initially design the cluster with multiple racks, even if the initial number of nodes might fit within one rack. Starting with multiple racks will enforce proper network topology and prevent future reconfiguration and hardware changes.

Also, as the number of nodes within the cluster increases, many of the tasks of managing the cluster also increase, such as updating node firmware or operating systems. Building a cluster management framework as part of the initial design and proactively considering the challenges of managing a large cluster will pay off significantly in the end.

Proactive planning for future scale out and the development of cluster management framework as a part of initial cluster design provides a foundation for future growth that will minimize hardware reconfigurations and cluster management issues as the cluster grows.

7.4 High availability considerations

When IBM Db2 Warehouse cluster is implemented, consider availability requirements as part of the final hardware and software configuration. IBM Db2 Warehouse cluster best practices provide significant protection against data loss. Generally, failures can be managed without causing an outage. There is redundancy that can be added to make a cluster even more reliable. Some consideration must be given to hardware and software redundancy.

An optional (but recommended) HAProxy load balancer can be set up on a separate node. This allows continued access to the cluster during a failover.

8 Acknowledgements

This reference architecture document has benefited from contributions and careful review comments provided by several colleagues. In particular, we gratefully acknowledge the collaboration and participation by Ajay Dholakia, Prasad Venkatachar, Russ Resnick and Ron Kunkel from Lenovo, Stewart Tate, Shubin Zhao, Bradley Yoo, Sven Oehme, Mitesh Shah and James Cho from IBM, and Raghu Moorthy and Ravikanth Durgavajhala from Intel.

Readers can contact Ajay Dholakia or Prasad Venkatachar from Lenovo for inquiries and information about the IBM Db2 Warehouse solution described in this reference architecture.

Resources

For more information, see the following resources:

Lenovo ThinkSystem SR650:

- Lenovo Press product guide: https://lenovopress.com/lp0644-lenovo-thinksystem-sr650-server

Lenovo RackSwitch G8052 (1GbE Switch):

- Lenovo Press product guide: https://lenovopress.com/tips1270-lenovo-rackswitch-g8052

Lenovo ThinkSystem NE10032 (40GbE/100GbE Switch):

- Lenovo Press product guide: https://lenovopress.com/lp0609-lenovo-thinksystem-ne10032-rackswitch

Lenovo XClarity Administrator:

- Lenovo Press product guide: https://lenovopress.com/tips1200-lenovo-xclarity-administrator

Lenovo Big Data Validated Design for Hortonworks Data Platform Using ThinkSystem Servers:

- Lenovo Press Solution page: https://lenovopress.com/lp0776

- IBM Db2 Warehouse
  - https://www.ibm.com/us-en/marketplace/db2-warehouse#product-header-top
  - https://www.ibm.com/support/knowledgecenter/en/SS6NHC/com.ibm.swg.im.dashdb.kc.doc/welcome.html
- IBM Spectrum Scale
  - https://www.ibm.com/support/knowledgecenter/STXKQY/ibmspectrumscale_welcome.html

Intel SSD products:

- https://www.intel.com/content/www/us/en/products/memory-storage/solid-state-drives/data-center-ssds.html

Document history

Version 1.0	March 16, 2018	• First version

Trademarks and special notices

References in this document to Lenovo products or services do not imply that Lenovo intends to make them available in every country.

Lenovo, the Lenovo logo, ThinkCentre, ThinkVision, ThinkVantage, ThinkPlus and Rescue and Recovery are trademarks of Lenovo.

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.

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

Intel, Intel Inside (logos), MMX, and Pentium are trademarks of Intel Corporation in the United States, other countries, or both.

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

Information is provided "AS IS" without warranty of any kind.

All customer examples described are presented as illustrations of how those customers have used Lenovo products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.

Information concerning non-Lenovo products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by Lenovo. Sources for non-Lenovo list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. Lenovo has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-Lenovo products. Questions on the capability of non-Lenovo products should be addressed to the supplier of those products.

All statements regarding Lenovo future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Contact your local Lenovo office or Lenovo authorized reseller for the full text of the specific Statement of Direction.

Some information addresses anticipated future capabilities. Such information is not intended as a definitive statement of a commitment to specific levels of performance, function or delivery schedules with respect to any future products. Such commitments are only made in Lenovo product announcements. The information is presented here to communicate Lenovo’s current investment and development activities as a good faith effort to help with our customers' future planning.

Performance is based on measurements and projections using standard Lenovo benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.

Photographs shown are of engineering prototypes. Changes may be incorporated in production models.

Any references in this information to non-Lenovo 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 Lenovo product and use of those websites is at your own risk.