Lenovo Big Data Validated Design for IBM SQL Analytics on ThinkSystem Servers

Last update: **26 January 2018**
Version 1.0

Describes the reference architecture for high performance SQL analytics with IBM Big SQL

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

Table of Contents

1 Introduction

This document describes the reference architecture for the Lenovo Big Data Validated Design for IBM SQL Analytics with ThinkSystem Servers. It provides a predefined and optimized hardware infrastructure for high performance SQL analytics using IBM Big SQL software running on Hortonworks Data Platform (HDP), a distribution of Apache Hadoop and Apache Spark software from Hortonworks. This reference architecture provides planning, design considerations, and best practices for implementing IBM Big SQL on HDP 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. With Apache Hadoop and Apache Spark emerging as popular big data storage and processing frameworks, enterprises are building so-called Data Lakes by employing these components. Running SQL analytics for Data Lakes and optimization of Enterprise Data Warehouse (EDW) modernization are areas of significant interest and focus. The IBM SQL Analytics solution described in this document is very well suited for implementing the infrastructure to support these modern analytics initiatives.

IBM Big SQL provides a flexible environment to implement SQL queries on Hadoop analytics. The users have the option to employ either the Spark SQL engine or the Big SQL engine for processing SQL queries on the data stored within Hadoop. Big SQL can transparently query data from both Hadoop and Relational Databases using the Fluidquery federation technology.

An SQL on Hadoop system must be balanced to deliver data 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 SQL on Hadoop 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 the SQL on Hadoop analytics 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 Hadoop, Spark and SQL concepts. For more information about Hadoop and Spark, see “Resources” on page 23.

2 Business problem and business value

This section describes business challenges faced by big data environments and the value provided by the IBM SQL Analytics 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.

Enterprises are incorporating large data lakes into their IT architecture to store all their data. The role of SQL in data analytics has been fundamental to the growth of analytics. A key challenge faced today by these enterprises is whether they need to adopt a totally different approach to analyze data stored in Data Lakes or can they continue to use the familiar SQL tools and techniques. The ability to effectively use of SQL over Hadoop with high performance is a key enabler for the adoption of big data in mainstream IT landscapes.

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. Until now, there was no effective way to harvest this opportunity. Today, IBM, Lenovo and partners use the latest big data technologies such as the in-memory processing capabilities of Spark in addition to the standard MapReduce scale-out capabilities of Hadoop to open the door to a world of new possibilities.

Hadoop is an open source software framework that is used to reliably manage and analyze large volumes of structured and unstructured data. Apache Spark enables the organizations to realize the real-time analytics and gain faster insights in today's competitive business environment. It provides significant performance advantages to analyze massive datasets with its in-memory processing engine and support for running standard SQL natively on an Apache Spark platform. IBM Big SQL is an enterprise grade SQL analytics engine optimized for delivering high performance at scale. This reference architecture document describes the performance and scalability benefits of deploying Apache Spark cluster on the Lenovo server platform 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 easliy
- 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:

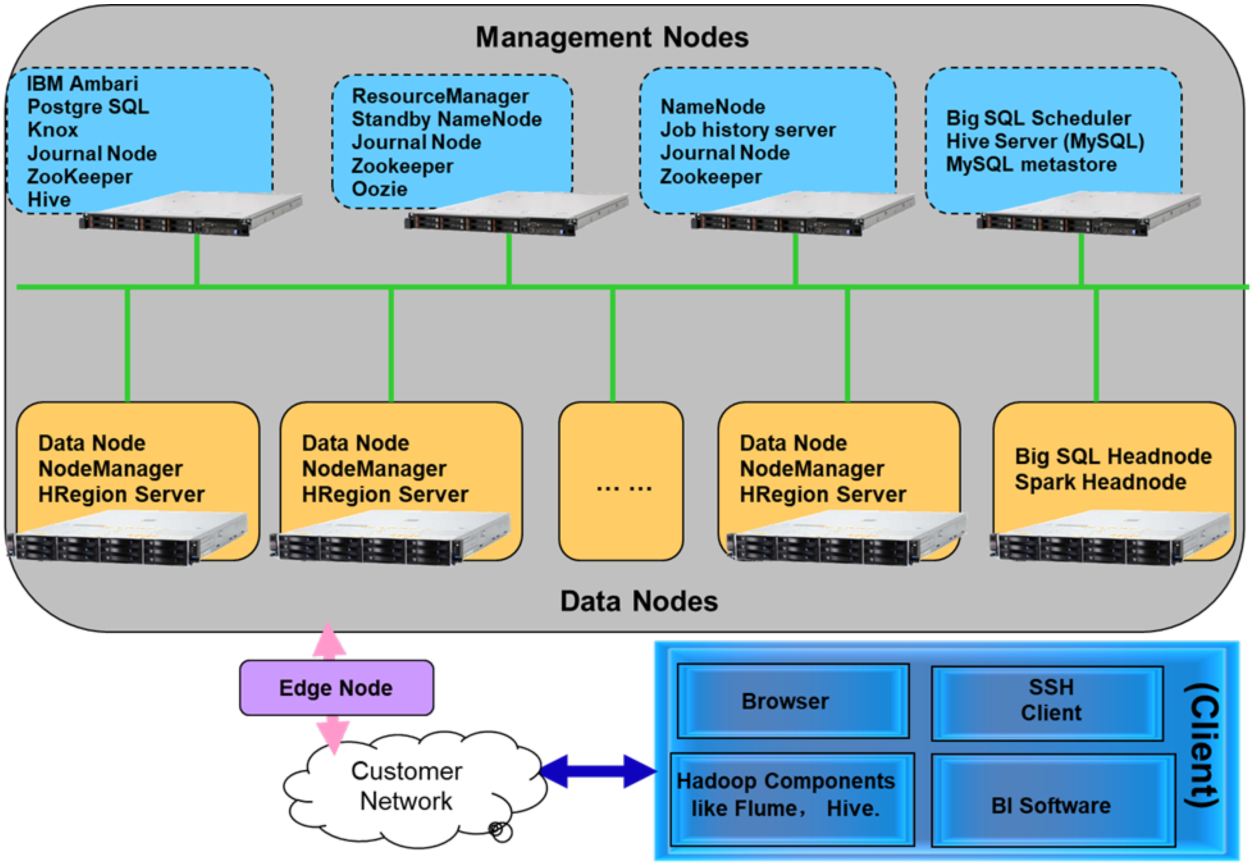
- Easy:
    - Ease of development
    - Easy management at scale
    - Advanced job management
    - Multi-tenancy
    - Easy to access data by various user types
- Dependable:
    - Data protection with snapshot and mirroring
    - Automated self-healing
    - Insight into software/hardware health and 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 SQL Analytics reference architecture solution has the following server roles:

- **Management nodes**: Nodes that are implemented on ThinkSystem SR630 servers. These nodes run services that are related to managing the cluster and coordinating the distributed environment.
- **Data nodes**: Nodes that are implemented on ThinkSystem SR650 servers. These nodes encompass daemons that are related to storing data and accomplishing work within the distributed environment.
- **Edge nodes**: Nodes that act as a boundary between the IBM SQL analytics cluster and the outside (client) environment.

Figure 1 shows the architecture overview of the IBM SQL Analytics reference architecture that uses Lenovo hardware.



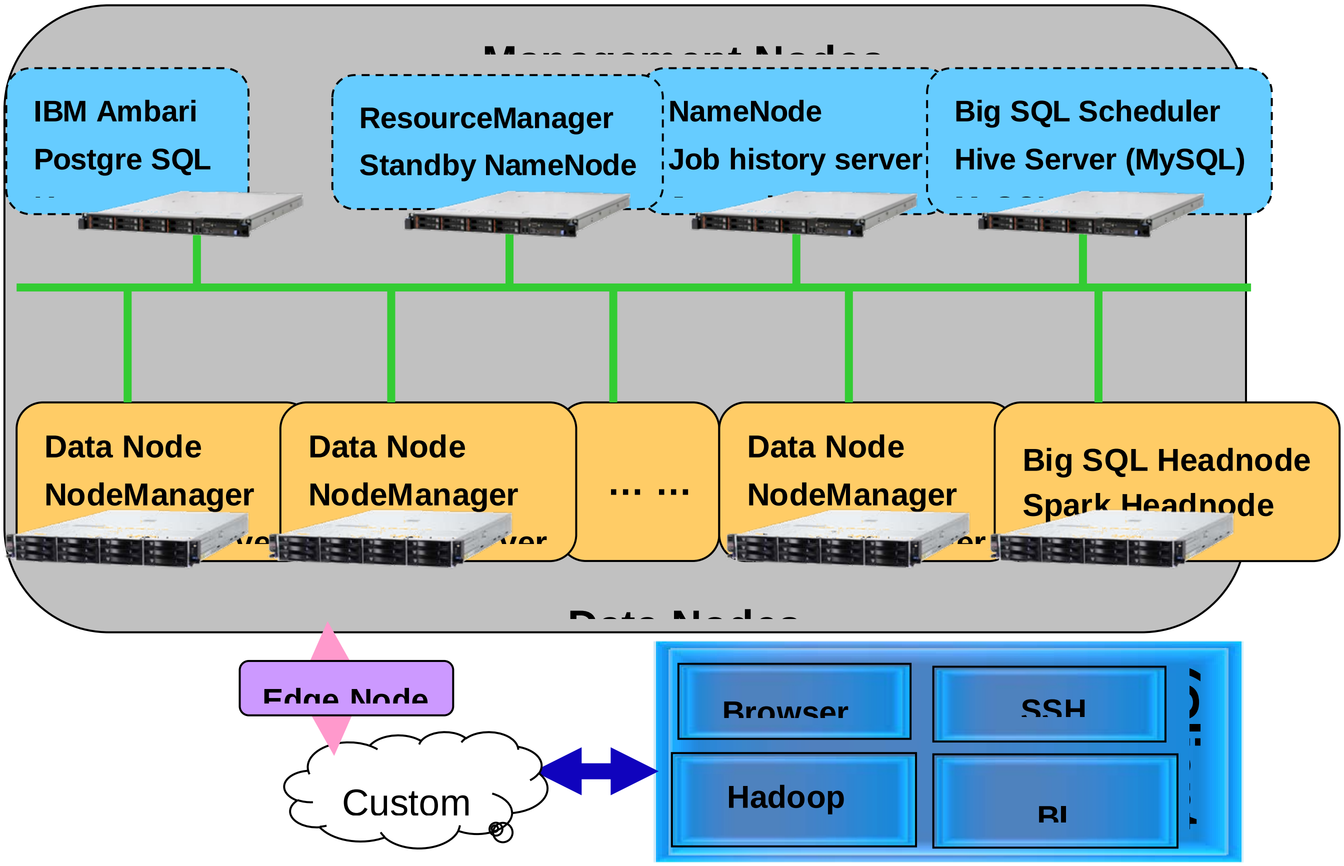


Figure 1. IBM SQL Analytics architecture overview

5 Component model

This section describes the high-level component model of the IBM SQL Analytics solution.

5.1 Hortonworks Data Platform overview

Hortonworks Data Platform is the industry's only truly secure, enterprise-ready, open source Apache Hadoop distribution based on a centralized architecture (YARN). It addresses the complete needs of “data-at-rest,” it powers real-time customer applications and it delivers robust analytics that accelerate decision-making and innovation.

The Hortonworks Data Platform for big data can be used for various use cases from batch applications that use MapReduce or Spark with data sources, such as click streams, to real-time applications that use sensor data.

Figure 2 shows the Hortonworks Hadoop collection of software frameworks, which make up the Hortonworks distribution of Apache Hadoop. Many of these Hadoop components are optional and provide specific functions to meet the requirements of customers.

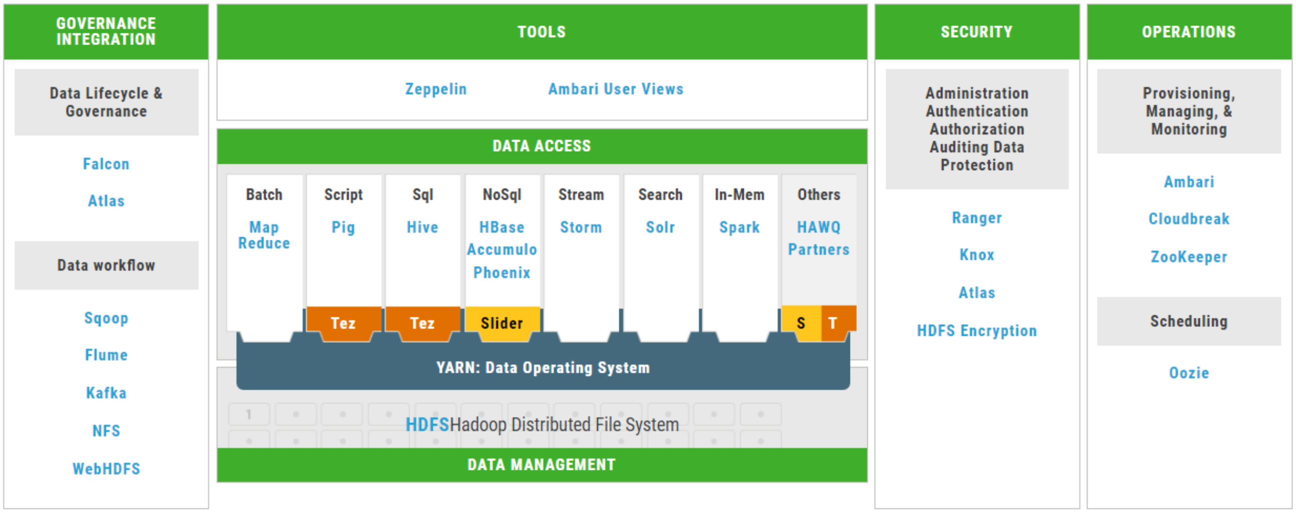


Figure 2 - Hortonworks Hadoop Collection of Software Frameworks

Hortonworks Data Platform contains the following components:

Data Management Components

YARN and Hadoop Distributed File System (HDFS) are the cornerstone components of Hortonworks Data Platform. While HDFS provides the scalable, fault-tolerant, cost-efficient storage for your big data lake, YARN provides the centralized architecture that enables you to process multiple workloads simultaneously. YARN provides the resource management and pluggable architecture for enabling a wide variety of data access methods.

Data Access Components

Hortonworks Data Platform includes a versatile range of processing engines that empower you to interact with the same data in multiple ways, at the same time. This means applications can interact with the data in the best way: from batch to interactive SQL or low latency access with NoSQL. Emerging use cases for data science, search and streaming are also supported with Apache Spark, Storm and Kafka. Other components include: Hive, Tez, Pig, Hbase and Accumulo.

Data Governance & Integration Components

HDP extends data access and management with powerful tools for data governance and integration. They provide a reliable, repeatable and simple framework for managing the flow of data in and out of Hadoop. This control structure, along with a set of tooling to ease and automate the application of schema or metadata on sources is critical for successful integration of Hadoop into your modern data architecture. The components include: Atlas, Falcon, Oozie, Scoop, Flume and Kafka.

Security Components

Security is woven and integrated into HDP in multiple layers. Critical features for authentication, authorization, accountability and data protection are in place to help secure HDP across these key requirements. Consistent with this approach throughout all of the enterprise Hadoop capabilities, HDP also ensures you can integrate and extend your current security solutions to provide a single, consistent, secure umbrella over your modern data architecture. These components include Knox, Ranger and Ranger KMS.

Operations Components

Operations teams deploy, monitor and manage a Hadoop cluster within their broader enterprise data ecosystem. Apache Ambari simplifies this experience. Ambari is an open source management platform for provisioning, managing, monitoring and securing the Hortonworks Data Platform. It enables Hadoop to fit seamlessly into your enterprise environment. These components include Ambari and Zookeeper.

Cloud Component

Cloudbreak, as part of Hortonworks Data Platform and powered by Apache Ambari, allows you to simplify the provisioning of clusters in any cloud environment including Amazon Web Services and Microsoft Azure. It optimizes your use of cloud resources as workloads change.

Spark

Apache Spark is a fast, in-memory data processing engine with elegant and expressive development APIs to allow data workers to efficiently execute streaming, machine learning or SQL workloads that require fast iterative access to datasets. With Spark running on Apache Hadoop YARN, developers everywhere can now create applications to exploit Spark's power, derive insights and enrich their data science workloads within a single, shared dataset in Hadoop.

The Hadoop YARN-based architecture provides the foundation that enables Spark and other applications to share a common cluster and dataset while ensuring consistent levels of service and response. Spark is now one of many data access engines that work with YARN in HDP.

Spark is designed for data science and its abstraction makes data science easier. Data scientists commonly use machine learning – a set of techniques and algorithms that can learn from data. These algorithms are often iterative, and Spark's ability to cache the dataset in memory greatly speeds up such iterative data processing, making Spark an ideal processing engine for implementing such algorithms.

For more information on all of the Hortonworks HDP Projects, see the following website:

http://hortonworks.com/apache/

The Hortonworks solution is operating system independent. Hortonworks HDP 2.6 supports many 64-bit Linux operating systems:

- Red Hat Enterprise Linux (RHEL), 64-bit
- CentOS, 64-bit
- Debian
- Oracle Linux, 64-bit
- SUSE Linux Enterprise Server (SLES), 64-bit
- Ubuntu, 64-bit

For more information about the versions of supported operating systems, see this website:

https://docs.hortonworks.com/HDPDocuments/Ambari-2.2.1.1/bk_Installing_HDP_AMB/content/_operating_systems_requirements.html

5.2 IBM Big SQL overview

With HDP as its foundation, IBM Big SQL enables users to query Hive and HBase data using ANSI-compliant SQL. While Hadoop is highly scalable, the Big SQL advanced cost-based optimizer and massively parallel processing (MPP) architecture can run queries smarter, not harder, supporting more concurrent users and more complex SQL with less hardware compared to other SQL solutions for Hadoop. Big SQL is also a powerful platform for data warehouse offloading and consolidation, a vital use case for many Hadoop users. Big SQL is the first and only SQL-on-Hadoop solution to understand commonly used SQL syntax from other vendors and products.

Hadoop has many next-generation analytics engines to solve big data problems. Big SQL has superior SQL-on-Hadoop performance using elastic boost technology. Big SQL also has deep Spark 2.1 integration, enabling new and varied use cases such as high-performance scans or high-performance inserts, updates or deletes. It also enables machine learning or graph analytics with Spark, with a single security model. Big SQL is a SQL engine for Hadoop that concurrently exploits Hive, HBase and Spark using a single database connection — even a single query. For this reason, Big SQL is also an effective hybrid engine.

6 Operational model

This section describes the hardware infrastructure aspects of the IBM SQL Analytics reference architecture. To support different customer environments, four different configurations are provided for supporting different amounts of data. Throughout the document, these configurations are referred to as starter rack, half rack, full rack, and multi-rack configuration sizes.

The discussion in the following sections provides details germane to the high-performance SQL analytics aspects of the design. For a general discussion of server hardware, cluster nodes, systems management, and networking infrastructure used in the design of the underlying Hadoop cluster, refer to

Lenovo Big Data Validated Design for HortonWorks Data Platform Using ThinkSystem Servers.

6.1 Overview

A typical deployment consists of cluster nodes, networking equipment, power distribution units, and racks.

- Management. The reference architecture handles both systems management (using Ambari) and hardware management (using Lenovo XClarity™ Administrator). 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" on page 23..
- Networking. The reference architecture specifies two networks: a data network and an administrative management network.
  - The data network creates a private cluster among multiple nodes and is used for high-speed data transfer across nodes, and for importing data into the IBM SQL Analytics cluster. The cluster typically connects to the customer's corporate data network. For cross-rack networking, additional switches per cluster are required.
  - The hardware management network is a 1 GbE network for out-of-band hardware management. Through the XClarity™ Controller management module (XCC) within the ThinkSystem SR650 server, the out-of-band network enables the hardware-level management of cluster nodes, such as node deployment, BIOS configuration, hardware failure status, and server power states.
- Predefined configurations. The predefined configurations can be implemented as-is or modified based on specific customer requirements, such as lower cost, improved performance, and increased reliability. Key workload requirements, such as the data growth rate, sizes of datasets, and data ingest patterns help in determining the proper configuration for a specific deployment. A best practice when an IBM SQL Analytics cluster infrastructure is designed is to conduct the proof of concept testing by using representative data and workloads to ensure that the proposed design works.

6.2 Cluster nodes

The IBM SQL Analytics reference architecture is implemented on a set of nodes that make up a cluster of data nodes, management nodes and optional edge nodes. Data nodes use ThinkSystem SR650 servers with locally attached storage while management nodes use ThinkSystem SR640 servers. Any edge nodes use the same server as management nodes. Data nodes run data (worker) services for storing and processing data.

Management nodes run management (control) services for coordinating and managing the cluster. Edge node (optional) acts as a boundary between the cluster and the outside (client) environment.

6.2.1 Management nodes

The Master node is the nucleus of the Hadoop Distributed File System (HDFS) and supports several other key functions that are needed on a Hortonworks cluster.

The Master node runs the following services:

YARN ResourceManager: Manages and arbitrates resources among all the applications in the system.

Hadoop NameNode: Controls the HDFS file system. The NameNode maintains the HDFS metadata, manages the directory tree of all files in the file system and tracks the location of the file data within the cluster. The NameNode does not store the data of these files.

ZooKeeper: Provides a distributed configuration service, a synchronization service and a name registry for distributed systems.

JournalNode: Collects, maintains and synchronize updates from NameNode.

HA ResourceManager: Standby ResourceManager that can be used to provide automated failover.

HA NameNode: Standby NameNode that can be used to provide automated failover.

Other non-master node services for Hadoop component management such as: Ambari server, HBase master, HiveServer2, and Spark History Server.

Table 2 lists the recommended components for a Master node and they can be customized according to client needs.

Table 1. Master node configuration

Component	Master node configuration
Server	ThinkSystem SR630
Processor	2x Intel® Xeon® Scalable Processors: 4114 Silver, 12-core 2.1Ghz
Memory - base	128 GB – 8x 16 GB 2666MHz RDIMM
Disk (OS / local storage)	OS: Dual M.2 128GB SSD with RAID1 Data: 4x 2TB 2.5” SAS HDD
HDD controller	ThinkSystem RAID 930-16i 4GB Flash 12Gb controller (with JBOD interface)
Hardware storage protection	RAID1: OS RAID10: NameNode/Metastore, Database, Zookeeper, QJN
Hardware management controller	Integrated XCLARITY™ CONTROLLER (XCC) with 1GBaseT dedicated interface or shared LAN interface
Data network adapter	ThinkSystem 10Gb 4-port SFP+ LOM

The Intel® Xeon® Scalable Processors and minimum memory specified in Table 2 is recommended to provide sufficient performance as a Hortonworks Master node. The M.2 SSD form factor is intended for Operating Storage in this reference architecture.

6.2.2 Data nodes

Table 3 lists the recommended system components for data nodes in this reference architecture.

Table 3. 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	1.5 TB: 24 x 64GB 2400MHz RDIMM
Disk (OS)	Dual M.2 128GB SSD with RAID1
Flash storage (data)	4x 2.0 TB Intel® SSD DC P4600 Series SSDs (AIC) 4x 2.0 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 3 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. Available data space assumes the use of Hadoop replication with three copies of the data (reduces effective disk space by 3x) plus a 25% reserve capacity so the HDFS file system is not constrained near term usage growth.

A minimum of three worker nodes are required as Hadoop has three copies of data by default. Three should be used for test or Proof of Concept (POC) environments only. A minimum of five worker nodes are required for production environment to reduce risk from losing more than one node at a time

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 a IBM SQL Analytics 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 as Hadoop has three copies of data by default. Three data nodes should be used for test or POC environments only. A minimum of five data nodes are required for production environment if there are data node failures.

6.2.3 Edge nodes (optional)

The optional edge node acts as a boundary between the IBM SQL Analytics 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 SR630 servers, other Lenovo servers, or other client-provided servers.

Although a IBM SQL Analytics cluster can have multiple edge nodes, depending on applications and workload, not every cluster rack needs to be connected to an edge node. However, every data node within the cluster must be a cluster data network IP address that is routable from within the corporate data network.

As gateways into the cluster, you must properly size edge nodes to ensure that they do not become a bottleneck for accessing the cluster, for example, during high volume ingest periods. Important: The number of edge nodes and the edge node server physical attributes that are required depend on ingest volume and velocity. Because of physical space constraints within a rack, adding an edge node to a rack can displace a data node. In low volume/velocity ingest situations (< 1 GB/hr), the Ambari management node can be used as an edge node. InfoSphere DataStage and InfoSphere Data Click servers can also function as edge nodes. When using InfoSphere DataStage or other ETL software, consult an appropriate ETL specialist for server selection. In Proof-of-Concept (PoC) situations, the edge node can be used to isolate both cluster networks (data and administrative/management) from the customer corporate network.

6.3 Networking

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

6.3.1 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.

6.3.2 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.

Hadoop has no dependency on the IMM2. 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. Administrators also can access all of the nodes in the cluster through the customer admin network, as shown in Figure 10 on page **Error! Bookmark not defined.**. 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.

6.4 Predefined cluster configurations

The intent of the predefined configurations is to ease initial sizing for customers and to show example starting points for three different-sized workloads.

- The half rack configuration consists of three data nodes, up to three management nodes and a pair of rack switches.
- The full rack configuration consists of 10 data nodes, up to three management nodes and a pair of rack switches.
- The multi-rack contains a total of 30 data nodes, up to 3 management nodes and a pair of switches.

The configuration is not limited to these sizes, and any number of data nodes is supported.

Table 3 lists four predefined configurations for the IBM SQL Analytics reference architecture. The table also lists the amount of space for data and the number of nodes that each predefined configuration provides. Storage space is described in two ways: the total amount of raw storage with 2 TB NVMe flash storage drives (raw storage) are used and the amount of space for the data that the customer has (available data space). Available data space assumes the use of Hadoop replication with three copies of the data and 25% capacity that is reserved for intermediate data (scratch storage). The estimates that are listed in Table 3 do not include extra space that is freed up by using compression because compression rates can vary widely based on file contents.

Table 2. Pre-defined configurations

	Starter	**Half rack**	**Full rack**	**Multi-rack**
Target dataset size	12 TB	25 TB	50 TB	100 TB
Number of Data Nodes	4	7	14	28
Storage capacity (raw)	64 TB	112 TB	224 TB	448 TB
Storage capacity (available data space)	16 TB	28 TB	56 TB	112 TB
Memory capacity	6.0 TB	10.5 TB	21 TB	42 TB
Number of Management Nodes	3	3	3	3
Number of Racks	1	1	1	2

Two types of rack layouts for the IBM SQL Analytics cluster are shown in Figure 13. Rack layout A shows a redundant design with 16 data nodes and 6 management nodes. Rack layout B shows a non-redundant design with 18 data nodes and 3 management nodes.

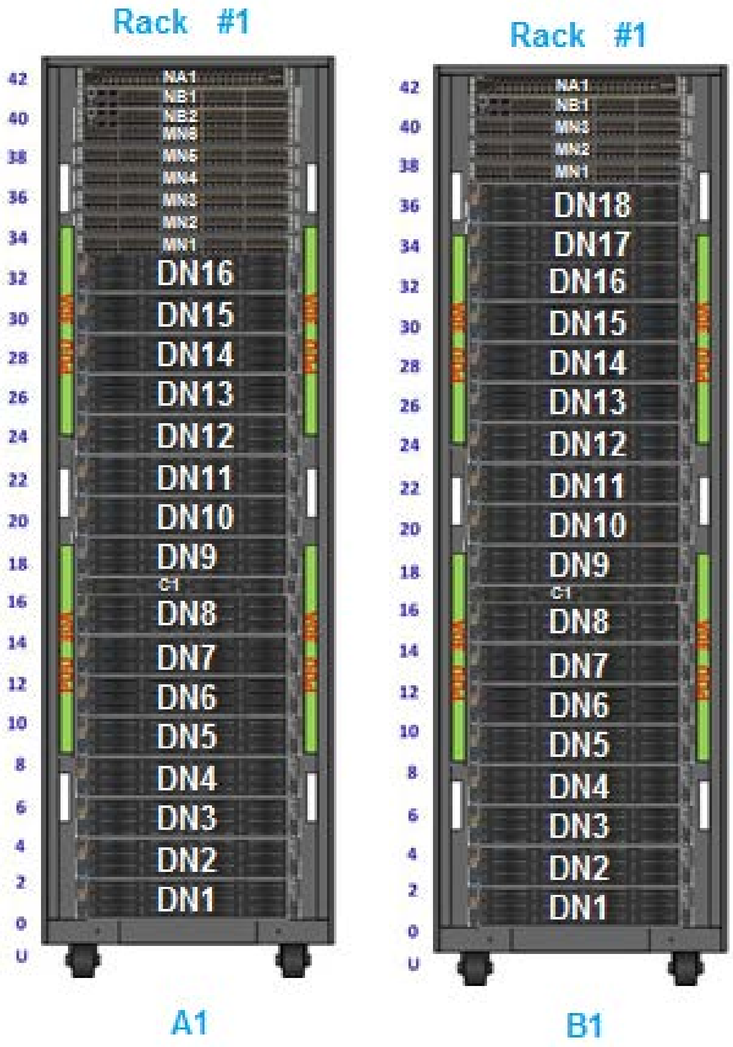


Figure 3. Rack layouts for IBM SQL Analytics cluster

7 Deployment considerations

This section describes various other considerations for deploying the IBM SQL Analytics 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 SQL Analytics 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 SQL Analytics cluster infrastructures.

7.1 Rack considerations

Within a rack, data nodes occupy 2U of space and management nodes and rack switches occupy 1U of space.

A one-rack IBM SQL Analytics implementation comes in three sizes: Starter rack, half rack, and full rack. These three sizes allow for easy ordering. However, reference architecture sizing is not rigid and supports any number of data nodes with the appropriate number of management nodes. Table 4 describes the node counts.

The IBM SQL Analytics cluster implementation can be deployed as a multi-rack solution. If the system is initially implemented as a multi-rack solution or if the system grows by adding more racks, to maximize fault tolerance, distribute the cluster management nodes across racks.

Table 3. Rack configuration node counts

Rack configuration size	Number of data nodes[1]	Number of management nodes[2]
Starter rack	3[3]	1 or 3
Half rack	9	1 or 3
Full rack with management nodes	18 [4]	1, 3, or 5
Full data node rack, no management nodes	20	0

In the reference architecture for the IBM SQL Analytics cluster, a fully populated predefined rack with two SN2700 switch and one G8052 switch can support up to 19 data nodes. However, the total number of data nodes that a rack can accommodate can vary based on the number of top-of-rack switches and management nodes that are required for the rack within the overall solution design. The number of data nodes can be calculated by the following equation:

Maximum number data nodes = (42U - (# 1U Switches + # 1U Management Nodes)) / 2

Edge nodes: This calculation does not consider edge nodes. Based on the client's choice of edge node, proportions can vary. Every two 1U edge nodes displace one data node, and every one 2U displaces one data node.

7.2 Designing for lower cost

There are several key modifications that can be made to lower the cost of a IBM SQL Analytics 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 IBMSQL Analytics 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

[1] Maximum number of data nodes per full rack based on network switches, management nodes, and data nodes. Adding edge nodes to the rack can displace additional data nodes.

[2] The number of management notes depends on development or the production/test environment type. For more information about selecting the correct number of management nodes, see "Management nodes" on page 13.

[3] The starter rack can be expanded to a full rack by adding more data and management nodes.

[4] A full rack with one management nodes can accommodate up to 19 data nodes.

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.

Table 5. Alternate data node types for IBM SQL Analytics cluster

Design Target	**CPU**	**Memory**	**Storage**
Highest performance	Xeon SP Platinum 8170	1.5 TB	16 TB (8x 2 TB NVMe)
Mid-level performance	Xeon SP Gold 6152	768 GB	6.4 TB (8x 800 GB NVMe)
Base-level performance	Xeon SP Gold 6140	384 GB	15.4 TB (16x 960 GB SSD)

7.3 Estimating storage capacity

When you are estimating storage space within a IBM SQL Analytics Hadoop cluster, consider the following points:

- For improved fault tolerance and performance, the Hadoop file system replicates data blocks across multiple cluster data nodes. By default, the file system maintains three replicas.
- Compression ratio is an important consideration in estimating disk space and can vary greatly based on file contents. If the customer's data compression ratio is unavailable, assume a compression ratio of 2.5:1.
- To ensure efficient file system operation and to allow time to add more storage capacity to the cluster if necessary, reserve 25% of the total capacity of the cluster.

Assuming the default three replicas maintained by the Hadoop file system, the raw data disk space, and the required number of nodes can be estimated by using the following equations:

*Total raw data disk space = (User data, uncompressed) * (4 / compression ratio)*

Total required data nodes = (Total raw data disk space) / (Raw data disk per node)

You should also consider future growth requirements when estimating disk space.

Based on these sizing principles, Table 8 on page 20 shows an example for a cluster that must store 250 TB of uncompressed user data. The example shows that the IBM SQL Analytics cluster needs 400 TB of raw disk to support 250 TB of uncompressed data. The 400 TB is for data storage and does not include operating system disk space. A total of 25 data nodes are required to support a deployment of this size.

Table 6. Example of storage sizing with 2TB NVMe solid-state drives

Description	**Value**
Size of uncompressed user data	250 TB
Compression ratio	2.5:1
Size of compressed data	100 TB
Storage multiplication factor	4
Raw data disk space needed for IBM SQL Analytics cluster	400 TB
Storage needed for Hadoop 3x replication	300 TB
Reserved storage for headroom	100 TB
Raw data disk per node (with 2 TB NVMe flash drives)	16 TB
Minimum number of nodes required (800/16)	25

Note: A 2.5:1compression ratio is an estimate based on measurements taken in a controlled environment. Compression results vary based on data and compression libraries used. Lenovo cannot guarantee compression results or compressed data storage amounts. Improved estimates can be calculated by testing customer data using appropriate compression libraries.

7.4 Scaling considerations

The Hadoop architecture is linearly scalable. 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 SQL Analytics solution reference architecture implementation, future scale out is a key consideration in the initial design. You must consider the two key 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 14 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.

7.5 High availability considerations

When IBM SQL Analytics cluster is implemented, consider availability requirements as part of the final hardware and software configuration. Typically, Hadoop is considered a *highly reliable* solution. Hadoop and IBM SQL Analytics 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.

7.5.1 Networking considerations

An active/active design is beyond the scope of what has been tested in this paper. An upcoming update will cover this. Some points that we will need to cover are.

1. Single or Dual Cards
2. Bonding Modes, Etherchannel, LACP, TLB, ALB
3. The hashing algorithm used with the bonding mode, Layer2, Layer2+Layer3, Layer3+Layer4
4. Inter switch communication links, are they needed, how many
5. Switch aggregation with technologies such as vPC, vLAG, mLAG

7.5.2 Hardware availability considerations

The redundancy of each individual data node is not necessary with IBM SQL Analytics cluster . Hadoop 3x replication provides built-in redundancy and makes loss of data unlikely. If Hadoop best practices are used, an outage from a data node loss is extremely unlikely as the workload can be dynamically re-allocated. The loss of a data node cannot stop workload processing; workload is automatically re-allocated to another data note.

Multiple management nodes are recommended so that if there is a failure, function can be moved to an operational management node. Having multiple management nodes does not resolve the issue of the NameNode being a single point of failure. For more information, see “Software availability considerations”.

Within racks, switches and nodes must have redundant power feeds with each power feed connected from a separate PDU.

7.5.3 Software availability considerations

Operating system availability is provided by using mirrored drives for the operating system.

NameNode HA is recommended and can be achieved by using three management nodes. Active and standby nodes communicate with a group of separate daemons called JournalNodes to keep their state synchronized. When any namespace modification is performed by the active NameNode, it durably logs a record of the modification to most of these JournalNodes. The standby NameNode can read the edits from the JournalNodes and is constantly watching them for changes to the edit log. As the standby Node sees the edits, it applies them to its own namespace.

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, Berni Schiefer and Jessi Chen from IBM, and Raghu Moorthy, Kshitij Doshi, Ravikanth Durgavajhala and Anil Patel from Intel.

Reader can contact Ajay Dholakia from Lenovo for inquiries and information about the IBM SQL Analytics solution described in this reference architecture.

Resources

For more information, see the following resources:

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

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

Lenovo ThinkSystem SR650 (Hortonworks Worker Node):

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

Lenovo ThinkSystem SR630 (Hortonworks Master node):

- Lenovo Press product guide: https://lenovopress.com/lp0643-lenovo-thinksystem-sr630-server

Lenovo RackSwitch G8052 (1GbE Switch):

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

Lenovo RackSwitch G8272 (10GbE Switch):

- Lenovo Press product guide: https://lenovopress.com/tips1267-lenovo-rackswitch-g8272

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

- IBM Big SQL
  - IBM Analytics Internet: https://www.ibm.com/us-en/marketplace/big-sql

Intel SSD products:

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

Hortonworks:

- Hortonworks Data Platform (HDP): http://hortonworks.com/products/data-center/hdp/

Document history

Version 1.0	January 23, 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.