ARCHITECTURE GUIDE

Hitachi Solution for Databases in an Enterprise Data Warehouse Offload Package for Oracle Database with MongoDB

Reference Architecture Guide

By Subhash Shinde, Shashikant Gaikwad

November 2019

Feedback

Hitachi Vantara welcomes your feedback. Please share your thoughts by sending an email message to SolutionLab@hitachivantara.com. To assist the routing of this message, use the paper number in the subject and the title of this white paper in the text.

Revision History

Revision	Changes	Date
MK-SL-177-00	Initial release	November 15, 2019

Table of Contents

Hitachi Solution for Databases in an Enterprise Data Warehouse Offload Package for Oracle Database with MongoDB

Reference Architecture Guide

Use this reference architecture guide to implement Hitachi Solution for Databases in an enterprise data warehouse offload package for Oracle Database. This Oracle converged infrastructure provides a high performance, integrated solution for business analytics using the following big data applications:

- Hitachi Advanced Server DS120
- Pentaho
- MongoDB

This architecture establishes best practices for environments where you can copy or move data in an enterprise data warehouse to a NoSQL database, such as MongoDB. You can then query your data from the NoSQL database instead of from the production Oracle database environment.

This reference architecture guide is for you if you are in one of the following roles and need to create a big data management solution:

- Data scientist
- Database administrator
- System administrator
- Storage administrator
- Database performance analyzer
- IT professional with the responsibility of planning and deploying an EDW offload solution

To use this reference architecture guide, you should have familiarity with the following:

- Hitachi Advanced Server DS220
- Hitachi Advanced Server DS120
- Pentaho
- MongoDB
- Oracle single instance database Release 18.3
- Big data and NoSQL
- IP networks
- Red Hat Enterprise Linux

Note — Testing of this configuration was done in a lab environment. Many things affect production environments beyond prediction or duplication in a lab environment. Follow the recommended practice of conducting proof-of-concept testing for acceptable results in a non-production, isolated test environment that otherwise matches your production environment before your production implementation of this solution.

Solution Overview

Use this reference architecture to implement Hitachi Solution for Databases in an enterprise data warehouse offload package for Oracle Database.

Business Benefits

This solution provides the following benefits:

- **Improve database manageability**

  You can take a "divide and conquer" approach to data management by moving data onto a lower cost storage tier without disrupting access to data.

- **Extreme scalability**

  Leveraging the extreme scalability of MongoDB, you can offload data from the Oracle servers onto commodity servers running big data solutions.

- **Lower total cost of ownership (TCO)**

  Reduce your capital expenditure by reducing the resources needed to run applications. Using MongoDB and low-cost storage makes it possible to keep information that is not deemed currently critical, but that you still might want to access, off the Oracle servers.

  This approach reduces the number of CPUs needed to run the Oracle database, optimizing your infrastructure. This potentially delivers hardware and software savings, including maintenance and support costs.

  Reduce the costs of running your workloads by leveraging less expensive, general purpose servers running MongoDB.

- **Improve availability**

  Reduce scheduled downtime by allowing database administrators to perform backup operations on a smaller subset of data. With offloading, perform daily backups on hot data, and less frequent backups on warm data.

  For an extremely large database, this can make the difference between having enough time to complete a backup in off-business hours or not.

  This also reduces the amount of time required to recover your data.

- **Analyze data without affecting the production environment**

  When processing and offloading data through Pentaho Data Integration, dashboards in Pentaho can analyze your data without affecting the performance of the Oracle production environment.

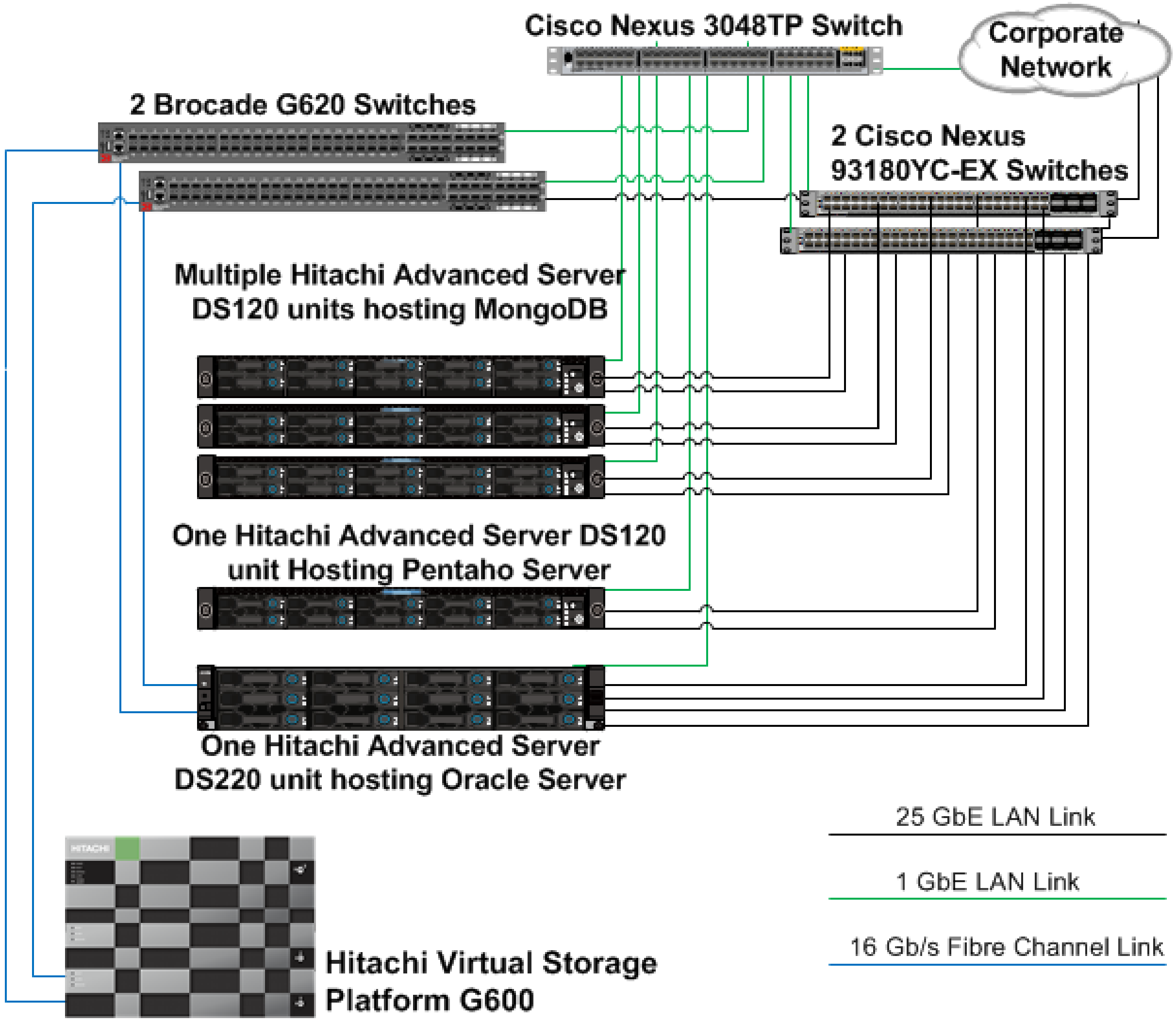
High Level Infrastructure

Figure 1 shows the high-level infrastructure for this solution.

This configuration of Hitachi Advanced Server DS120 provides the following characteristics:

- Fully redundant hardware
- High compute and storage density
- Flexible and scalable I/O options
- Sophisticated power and thermal design to avoid unnecessary operation expenditures
- Quick deployment and maintenance

Figure 1


Cisco Nexus 3048TP Switch
Corporate Network
2 Brocade G620 Switches
2 Cisco Nexus 93180YC-EX Switches
Multiple Hitachi Advanced Server DS120 units hosting MongoDB
One Hitachi Advanced Server DS120 unit Hosting Pentaho Server
One Hitachi Advanced Server DS220 unit hosting Oracle Server
25 GbE LAN Link
1 GbE LAN Link
16 Gb/s Fibre Channel Link
Hitachi Virtual Storage Platform G600

Note — Although the test environment used a single server for the Oracle environment, an enterprise data warehouse is likely to have a two- or more node configuration for the Oracle environment. This solution gives flexibility to use an existing Oracle environment, as well.

To avoid any performance impact to the production database, Hitachi Vantara recommends using a configuration with a dedicated IP network for the following:

- Production Oracle database
- Pentaho server
- MongoDB servers

Uplink speed to the corporate network depends on your environment and requirements. The Cisco Nexus 93180YC-EX switches can support uplink speeds of 40 GbE or 100 GbE, if higher bandwidth is required.

For validation testing, this solution used Hitachi Unified Compute Platform CI in a solution for an Oracle Database architecture with Hitachi Advanced Server DS220, Hitachi Virtual Storage Platform G600, and two Brocade G620 SAN switches hosting Oracle enterprise data warehouse. You can use your existing Oracle database environment, purchase Hitachi Unified Compute Platform CI to host Oracle Real Application Clusters or use a standalone solution to host the enterprise data warehouse.

Key Solution Components

The key solution components for this solution are listed in Table 1, “Hardware Components,” on page 4 and Table 2, “Software Components,” on page 6.

TABLE 1. HARDWARE COMPONENTS

Hardware	Detailed Description	Firmware or Driver Version	Quantity
Hitachi Virtual Storage Platform G600 (VSP G600)	■ One controller ■ 8 × 16 Gb/s Fiber Channel ports ■ 8 × 12 Gb/s backend SAS ports ■ 256 GB cache memory ■ 40 × 960 GB SSDs, plus 2 spares ■ 16 Gb/s × 2 ports CHB	83-05-29-40/00	1

TABLE 1. HARDWARE COMPONENTS (CONTINUED)

Hardware	Detailed Description	Firmware or Driver Version	Quantity
Hitachi Advanced Server DS220 (Oracle host)	2 Intel Xeon Gold 6140 CPU @ 2.30 GHz 768 GB (64GB × 12) DIMM DDR4 synchronous registered (buffered) 2666 MHz 1 x 3516 RAID controller	BIOS: S5BH3B14.H01 BMC: 4.62.06	1
	Intel XXV710 Dual Port 25 GbE NIC cards	6.128(6.80)	2
	Emulex Light Pulse LPe31002-M6 2-Port 16 Gb/s Fibre Channel adapter	12.0.193.13	2
	1.2 TB SAS HDD		2
Hitachi Advanced Server DS120 (MongoDB or Hitachi Content Platform host)	2 Intel Xeon Silver 4110 CPU @ 2.10 GHz 2 × 128 GB MLC SATADOM for boot 384 GB (32 GB × 12) DIMM DDR4 synchronous registered (buffered) 2666 MHz 1 x 3516 RAID controller	BIOS: S5BH3B14.H01 BMC: 4.62.06	3
	Intel XXV710 Dual Port 25 GbE NIC cards	6.128(6.80)	6
	1.2 TB SAS HDD		4
	960 GB SSD SATA (6.0 Gb/s)		4
Hitachi Advanced Server DS120 (Pentaho host)	2 Intel Xeon Silver 4110 CPU @ 2.10 GHz 2 × 128 GB MLC SATADOM for boot 128 GB (32 GB × 4) DIMM DDR4 synchronous registered (buffered) 2666 MHz 1 x 3516 RAID controller	BIOS: S5BH3B14.H01 BMC: 4.62.06	1
	1.2 TB SAS HDD		2
	Intel XXV710 Dual Port 25 GbE	6.128(6.80)	2
Brocade G620 switches	48 port Fiber Channel switch 16 Gb/s SFPs Brocade hot-pluggable SFP+, LC connector	V8.0.1	2

TABLE 1. HARDWARE COMPONENTS (CONTINUED)

Hardware	Detailed Description	Firmware or Driver Version	Quantity
Cisco Nexus 93180YC-EX switches	■ 48 × 10/25 GbE Fiber Channel ports ■ 6 × 40/100 Gb/s quad SFP (QSFP28) ports	■ 7.0(3)I4(7)	2
Cisco Nexus 3048TP switch	■ 1 GbE 48-Port Ethernet switch	■ 7.0(3)I4(7)	1

TABLE 2. SOFTWARE COMPONENTS

Software	Version	Function
Red Hat Enterprise Linux	Version 7.6 Kernel Version: 3.10.0-957.el7.x86_64	Operating system
Oracle	18.3.0.0.0	Database software
Pentaho Data Integration	8.3 (release 8.3.0.0-371)	Extract-transfer-load software
MongoDB	4.2	NoSQL database
Red Hat Enterprise Linux Device Mapper Multipath	0.4.9-127	Multipath Software
Hitachi Storage Navigator [Note 1]	Microcode dependent	Storage management software
Hitachi Storage Advisor (HSA) [Note 1]	3.3	Storage orchestration software

[Note 1] These software programs were used for this Oracle Database architecture built on Hitachi Unified Compute Platform CI to validate this solution. They may not be required for your implementation.

This solution was tested with Pentaho Data Integration version 8.2 and version 8.3 as well.

Note — At the time of publishing, the Hitachi Solutions for Analytics Infrastructure OS Installer only supports RHEL 7.3 and 7.4 automated install. Although the testing in the lab for this release was done with RHEL 7.6, customers may still use RHEL 7.4 as a base install prior to updating to 7.6 and installing their NoSQL database and Pentaho application.

Pentaho

A unified data integration and analytics program, Pentaho addresses the barriers that block your organization's ability to get value from all your data. Simplify preparing and blending any data with a spectrum of tools to analyze, visualize, explore, report, and predict. Open, embeddable, and extensible, Pentaho ensures that each member of your team — from developers to business users — can translate data into value.

- **Internet of things** — Integrate machine data with other data for better outcomes.
- **Big data** — Accelerate value with Apache Hadoop, NoSQL, and other big data programs.
- **Data integration** — Access, manage, and blend any data from any source.

  This solution uses “Pentaho Data Integration” on page 15 to drive the extract, transform, and load (ETL) process. The end target of this process is a MongoDB database.
- **Business analytics** — Turn data into insights with embeddable analytics.

For Pentaho 8.3, read what features are available.

Hitachi Advanced Server DS120

Optimized for performance, high density, and power efficiency in a dual-processor server, Hitachi Advanced Server DS120 delivers a balance of compute and storage capacity. This rack mounted server has the flexibility to power a wide range of solutions and applications.

The highly scalable memory supports up to 3 TB using 24 slots of 2666 MHz DDR4 RDMM. DS120 is powered by the Intel Xeon scalable processor family for complex and demanding workloads. There are flexible OCP and PCIe I/O expansion card options available. This server supports up to 12 storage devices with up to 4 NVMe drives.

Hitachi Advanced Server DS220

With a combination of two Intel Xeon Scalable processors and high storage capacity in a 2U rack-space package, Hitachi Advanced Server DS220 delivers the storage and I/O to meet the needs of converged solutions and high-performance applications in the data center.

The Intel Xeon Scalable processor family is optimized to address the growing demands on today’s IT infrastructure. The server provides 24 slots for high-speed DDR4 memory, allowing up to 3 TB of memory per node when 238 GB DIMMs are used.

Hitachi Virtual Storage Platform G Series Family

The Hitachi Virtual Storage Platform G series family enables the seamless automation of the data center. It has a broad range of efficiency technologies that deliver maximum value while making ongoing costs more predictable. You can focus on strategic projects and to consolidate more workloads while using a wide range of media choices.

The benefits start with Hitachi Storage Virtualization Operating System RF. This includes an all new enhanced software stack that offers up to three times greater performance than our previous midrange models, even as data scales to petabytes.

Virtual Storage Platform G series offers support for containers to accelerate cloud-native application development. Provision storage in seconds, and provide persistent data availability, all the while being orchestrated by industry leading container platforms. Moved these workloads into an enterprise production environment seamlessly, saving money while reducing support and management costs.

This solution was validated with Virtual Storage Platform G600, which supports Oracle Real Application Clusters. You may use any Virtual Storage Platform F series (on page 8) or VSP G series product.

Hitachi Virtual Storage Platform F Series Family

Use Hitachi Virtual Storage Platform F series family storage for a flash-powered cloud platform for your mission critical applications. This storage meets demanding performance and uptime business needs. Extremely scalable, its 4.8 million random read IOPS allows you to consolidate more applications for more cost savings.

Hitachi Storage Virtualization Operating System RF is at the heart of the Virtual Storage Platform F series family. It provides storage virtualization, high availability, flash optimized performance, quality of service controls, and advanced data protection. This proven, mature software provides common features, management, and interoperability across the Hitachi portfolio. This means you can reduce migration efforts, consolidate assets, reclaim space, and extend life.

Reduce risks and solve problems faster. Integrated power analytics and automation features bring artificial intelligence to your data center. Cloud-assessible monitoring tools give your product support experts access wherever they have an internet connection for fast troubleshooting and remediation.

Brocade Switches

Brocade and Hitachi Vantara partner to deliver storage networking and data center solutions. These solutions reduce complexity and cost, as well as enable virtualization and cloud computing to increase business agility.

Optionally, this solution uses the following Brocade product:

- **Brocade G620 switch, 48-port Fibre Channel**

  In this solution, the SAN switches are optional. You may use direct connect under certain circumstances. Check the support matrix to ensure support for your choice.

Cisco Nexus Data Center Switches

Cisco Nexus data center switches are built for scale, industry-leading automation, programmability, and real-time visibility.

This solution uses the following Cisco switches to provide Ethernet connectivity:

- Nexus 93180YC-EX, 48-port 10/25 GbE switch
- Nexus 3048TP, 48-port 1GbE Switch

MongoDB

MongoDB is a document database with the scalability and flexibility that you want with the querying and indexing that you need. MongoDB’s document model is simple for developers to learn and use, while still providing all the capabilities needed to meet the most complex requirements at any scale.

Oracle Database

Oracle Database has a multi-tenant architecture so you can consolidate many databases quickly and manage them as a cloud service. Oracle Database also includes in-memory data processing capabilities for analytical performance. Additional database innovations deliver efficiency, performance, security, and availability. Oracle Database comes in two editions: Enterprise Edition and Standard Edition 2.

Oracle Automatic Storage Management (Oracle ASM) is a volume manager and file system for Oracle database files. This supports single-instance Oracle Database and Oracle Real Application Clusters configurations. Oracle ASM is the recommended storage management solution that provides an alternative to conventional volume managers, file systems, and raw devices.

Red Hat Enterprise Linux

Red Hat Enterprise Linux delivers military-grade security, 99.999% uptime, support for business-critical workloads, and so much more. Ultimately, the platform helps you reallocate resources from maintaining the status quo to tackling new challenges.

Device mapper multipathing (DM-Multipath) allows you to configure multiple I/O paths between server nodes and storage arrays into a single device.

These I/O paths are physical SAN connections that can include separate cables, switches, and controllers. Multipathing aggregates the I/O paths, creating a new device that consists of the aggregated paths.

Note — At the time of publishing, the Hitachi Solutions for Analytics Infrastructure Operating System Installer only supports Red Hat Enterprise Linux v7.3 and v7.4 in an automated installation. Although the testing in the lab for this release was done with Red Hat Enterprise Linux 7.6, you may still install Red Hat Enterprise Linux v7.4 as a base prior to updating to version 7.6 and installing Pentaho.

Solution Design

This describes the reference architecture environment to implement Hitachi Solution for the Databases in an enterprise data warehouse offload package for Oracle Database. The environment uses Hitachi Advanced Server DS120 and Hitachi Advanced Server DS220.

The infrastructure configuration includes the following:

- **Pentaho server**

  There is one Hitachi Advanced Server DS120 configured to run Pentaho.

- **MongoDB servers**

  There are at least three servers configured to run MongoDB. The number of MongoDB servers can be expended, based on size of working set, sharding, and other factors. This solution uses local SSDs with RAID-10 protection for the MongoDB hosts.

- **IP network connection**

  There are IP connections to connect the Pentaho server, the MongoDB servers, and the Oracle server through Cisco Nexus switches

- **Oracle Database Architecture**

  This infrastructure hosts Oracle Enterprise Data Warehouse. The tested infrastructure included the following built on this Hitachi Unified Compute Platform CI configuration:

  - One Hitachi Advanced Server DS220
  - Two Brocade G620 SAN switches
  - Hitachi Virtual Storage Platform G600

  In your implementation, any Virtual Storage Platform F series or Virtual Storage Platform G series model can be used.

Server and Application Architecture

This reference architecture uses the following:

- Three Hitachi Advanced Server DS120 to host MongoDB host configuration
- One Advanced Server DS120 to host Pentaho.

The architecture provides the compute power for the MongoDB database to handle complex database queries and a large volume of transaction processing in parallel.

Table 3 describes example details of a server configuration for this solution.

TABLE 3. HITACHI ADVANCED SERVER DS120 SPECIFICATIONS

Server	Server Name	Role	CPU Cores	RAM
MongoDB Server 1	mongodbnode1	MongoDB node 1	16	384 GB (32 GB × 12)
MongoDB Server 2	mongodbnode2	MongoDB node 2	16	384 GB (32 GB × 12)
MongoDB Server 3	mongodbnode3	MongoDB node 3	16	384 GB (32 GB × 12)
Pentaho Server	Pentaho	Pentaho Server	16	128 GB (32 GB × 4)

Table 4 shows the server BIOS and Red Hat Enterprise Linux 7.6 kernel parameters for hosting MongoDB.

TABLE 4. BIOS AND RED HAT ENTERPRISE LINUX 7.6 KERNEL PARAMETERS FOR HOSTING MONGODB

Parameter Category	Setting	Value
BIOS	NUMA	DISABLE
	DISK READ AHEAD	DISABLE
RHEL 7.6 Kernel	ulimit	64000
	vm.dirty_ratio	15
	vm.dirty_background_ratio	5
	vm.swappiness	1
	tansparent_hugepage	never
	IO scheduler	noop

Table 5 shows the server BIOS and Red Hat Enterprise Linux 7.6 kernel parameters for hosting Pentaho.

TABLE 5. BIOS AND RED HAT ENTERPRISE LINUX 7.6 KERNEL PARAMETERS FOR HOSTING PENTAHO

Parameter Category	Setting	Value
BIOS	NUMA	ENABLE
	DISK READ AHEAD	DISABLE
Red Hat Enterprise Linux 7.6 Kernel	ulimit	64000
	vm.dirty_ratio	15
	vm.dirty_background_ratio	5
	vm.swappiness	1
	tansparent_hugepage	never
	IO scheduler	noop

For a Pentaho server running on Microsoft Windows Server, read Increase the Pentaho Server memory limit.

Storage Architecture

This describes the storage architecture for this solution.

This solution was validated with Hitachi Virtual Storage Platform G600, which supports Oracle Real Application Clusters. You may use any Virtual Storage Platform F series or VSP G series product in your implementation of this environment.

Storage Configuration for MongoDB

Configure one RAID-10 group, or a total of four SSDs, on each MongoDB server. This uses recommended practices with Hitachi Advanced Server DS120 and MongoDB for the design and deployment of storage for MongoDB. For the best performance and size to accommodate the Oracle Enterprise Data Warehouse offloading space, adjust the number of RAID groups and/or size of SSDs to meet your business requirements.

Table 6, “Sample Storage Configuration for MongoDB,” on page 12 shows a sample storage configuration for MongoDB in this solution.

TABLE 6. SAMPLE STORAGE CONFIGURATION FOR MONGODB

RAID Level	RAID-10
Drive Type	960 GB SSDs
Number of Drives	4
Virtual Volume size	1.745 GB
Number of Virtual Volumes	1
Total Useable Capacity	1.745 TB
File System Type	XFS
File System Block Size	4 KB
Disk Readahead (BIOS Setting)	Disable

File System Considerations for MongoDB Database

Hitachi Vantara recommends using XFS with a default block size of 4 KB for the MongoDB database, as XFS generally performs better with MongoDB. This default block size provides balanced performance for mixed workloads.

In addition, you can use Red Hat Enterprise Linux 7.6 native file system EXT4. However, the EXT3 file system should be avoid due to its poor pre-allocation performance.

With the WiredTiger storage engine, XFS is strongly recommended to avoid performance issues that may occur when using EXT4, as Production Notes on the MongoDB website indicates.

For a shared MongoDB database environment, the file system configuration can be different to accommodate several types of workloads. The components in the solution set in this reference architecture have the flexibility for use in various deployment scenarios to provide the right balance between performance and ease of management for a given scenario.

System Management Server

This reference architecture uses a shared Hitachi Advanced Server DS120 server for the VMware ESXi management server configuration. Configure virtual machines to run Hitachi Storage Advisor.

Network Architecture

This architecture requires the following separate networks for the Pentaho server and the MongoDB servers:

- **Public Network** — This network must be scalable. In addition, it must meet the low latency needs of the network traffic generated by the servers running applications in the environment.
- **Management Network** — This network provides BMC connections to the physical servers.

Hitachi Vantara recommends using pairs of 25 Gb/s NICs for the public network and 1 Gb/s LOM for the management network.

Observe these points when configuring a public network in your environment:

- For each server in the configuration, use at least two identical, high-bandwidth, low-latency NICs for the public network.
- Use NIC bonding to provide failover and load balancing within a server. If using two dual-port NICs, NIC bonding can be configured across two cards.
- Set all NICs to full duplex mode.

Physical Network Configuration

Figure 2 shows the network configuration in this solution.

Figure 2

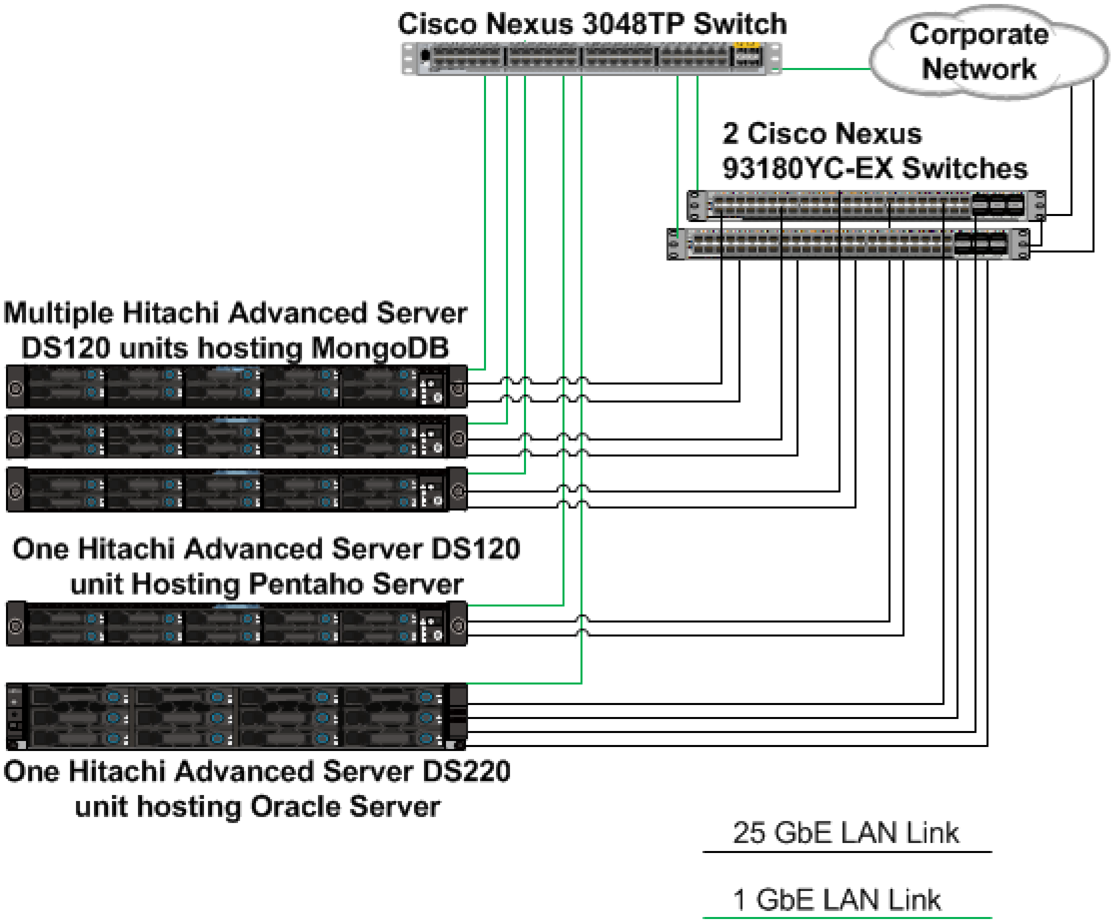

Cisco Nexus 3048TP Switch
Corporate Network
2 Cisco Nexus 93180YC-EX Switches
Multiple Hitachi Advanced Server DS120 units hosting MongoDB
One Hitachi Advanced Server DS120 unit Hosting Pentaho Server
One Hitachi Advanced Server DS220 unit hosting Oracle Server
25 GbE LAN Link
1 GbE LAN Link

Table 7 shows the network configuration, IP addresses, and name configuration that was used when testing the environment with Hitachi Advanced Server DS120 and Hitachi Advanced Server DS220. Your implementation of this solution can differ.

Configure pairs of ports from different physical NIC cards to avoid a single point of failure when installing two NICs on each server. However, this environment supports using one NIC on the MongoDB server or servers and the Pentaho server for lower cost.

TABLE 7. NETWORK CONFIGURATION AND IP ADDRESSES FOR HITACHI ADVANCED SERVER DS120 AND ADVANCED SERVER DS220

Server	NIC Ports	Subnet	NIC BOND	IP Address	Network	Bandwidth (Gb/s)	Cisco Nexus Switch	
							Switch Number	Port
DS120 Server 1 (MongoDB)	NIC - 0	242	Bond0	10.x.x.21	Public	25	1	11
	NIC - 3					25	2	
	BMC- Dedicated NIC	208	-	10.x.x.65	Management	1	3	11
DS120 Server 2 (MongoDB)	NIC - 0	242	Bond0	10.x.x.22	Public	25	1	12
	NIC - 3					25	2	
	BMC- Dedicated NIC	208	-	10.x.x.66	Management	1	3	12
DS120 Server 3 (MongoDB)	NIC - 0	242	Bond0	10.x.x.23	Public	25	1	13
	NIC - 3					25	2	
	BMC- Dedicated NIC	208	-	10.x.x.67	Management	1	3	13
DS120 Server 4 (Pentaho)	NIC - 0	242	Bond0	10.x.x.24	Public	25	1	14
	NIC - 3					25	2	
	BMC- Dedicated NIC	208	-	10.x.x.64	Management	1	3	14
DS220 Server (Oracle)	NIC-0	242	Bond0	10.x.x.25	Public	25	1	15
	NIC-3					25	2	
	BMC- Dedicated NIC	208	-	10.x.x.68	Management	1	3	15

Data Analytics and Performance Monitoring

Use this for data analytics and performance monitoring with this solution.

Hitachi Storage Advisor

By reducing storage infrastructure management complexities, Hitachi Storage Advisor simplifies management operations. This helps you to rapidly configure storage systems and IT services for new business applications.

Enterprise Data Offload Workflow

Use Hitachi's Global Services engagement to automate the Oracle Enterprise Data Workflow mapping of large Oracle data sets to MongoDB, and then offload the data using Pentaho Data Integration. The engagement leverages a lab tested automation tool that creates a transformation Kettle file with Spoon for the data offload.

This auto-generated transformation transfers row data from Oracle database tables or views in a schema to MongoDB. Pentaho Data Integration uses this transformation directly.

Pentaho Data Integration

Pentaho Data Integration allows you to ingest, blend, cleanse, and prepare diverse data from any source. With visual tools to eliminate coding and complexity, Pentaho puts all data sources and the best quality data at the fingertips of businesses and IT users.

Using intuitive drag-and-drop data integration coupled with data agnostic connectivity, your use of Pentaho Data Integration can span from flat files and RDBMS to Apache Hadoop and beyond. Go beyond a standard extract-transform-load (ETL) designer to scalable and flexible management for end-to-end data flows.

In this reference architecture, the end target of the ETL process is a MongoDB database.

Engagement Led Toolkit for Automatic Enterprise Data Workflow Offload

You can engage Hitachi Global Services to use the toolkit to automate the configuration of an Oracle Enterprise Data Workflow that uses existing user accounts with appropriate permissions for Oracle and MongoDB database access. The toolkit validates the Oracle and MongoDB connections first and then proceed with further options.

The main options to generate a Pentaho Data Integration transformation are the following:

- Transfer the entire Oracle schema to MongoDB
- Transfer Oracle tables based on partition to MongoDB
- Transfer specific Oracle table rows based on date range
- Transfer specific Oracle table rows based on column key value

Example Workflow Offloads Using Pentaho Data Integration

These examples are used by Pentaho Data Integration for the enterprise data warehouse offload. These are created using the graphical user interface in Pentaho Data Integration.

Full Table Data Copy from Oracle to MongoDB

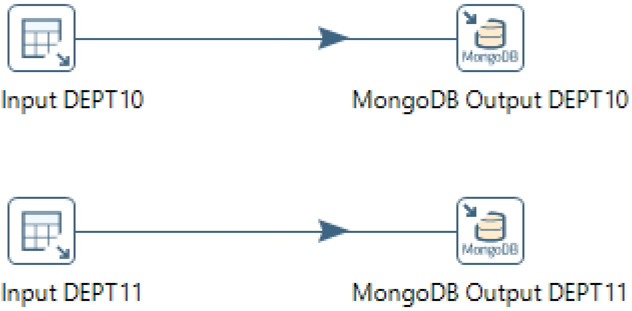
It is a challenge to convert data types between two database systems. With the graphical user interface in Pentaho Data Integration, you can do data type conversion with a few clicks. No coding is needed. This example demonstrates how to convert an unsupported time stamp into a string before being copied to the MongoDB database.

Use the graphical user interface in Pentaho Data Integration to construct a workflow to copy all the data from an Oracle table to MongoDB. This example uses database “pentaho71” and collection “edw11.”

User can define data connections for the Pentaho server so Pentaho Data Integration can access data from sources like Oracle Database. See Define Data Connections for the Pentaho Server for these procedures.

Figure 3 shows the Pentaho Data Integration workflow for a full table data copy from Oracle to MongoDB in the user interface.

Figure 3



To create a full table data copy workflow, do the following.

1. To make the settings for this workflow, double-click **Table Input DEPT10** (Figure 3). The **Table input** dialog box opens (Figure 4). Make the settings to input the table and click **OK**.

Figure 4

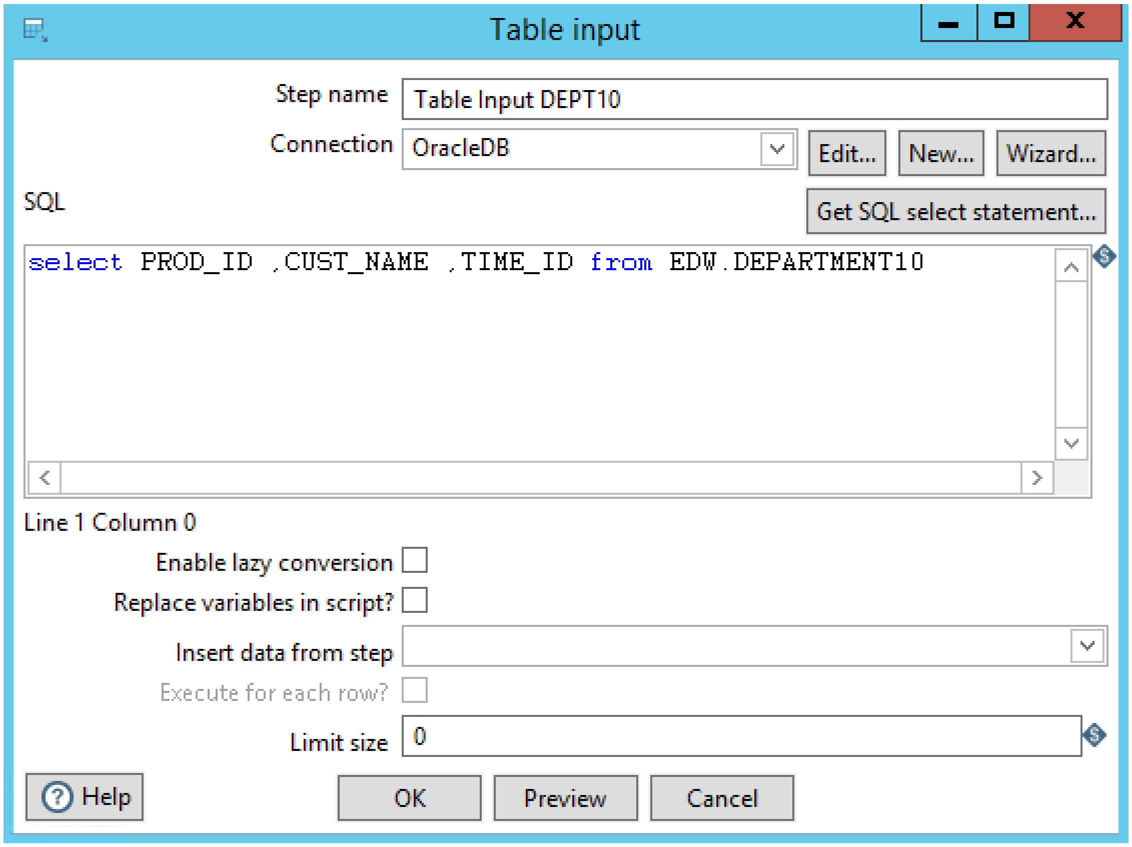


Table input

Step name: Table Input DEPT10

Connection: OracleDB

SQL

```
select PROD_ID ,CUST_NAME ,TIME_ID from EDW.DEPARTMENT10
```

Line 1 Column 0

Enable lazy conversion □

Replace variables in script? □

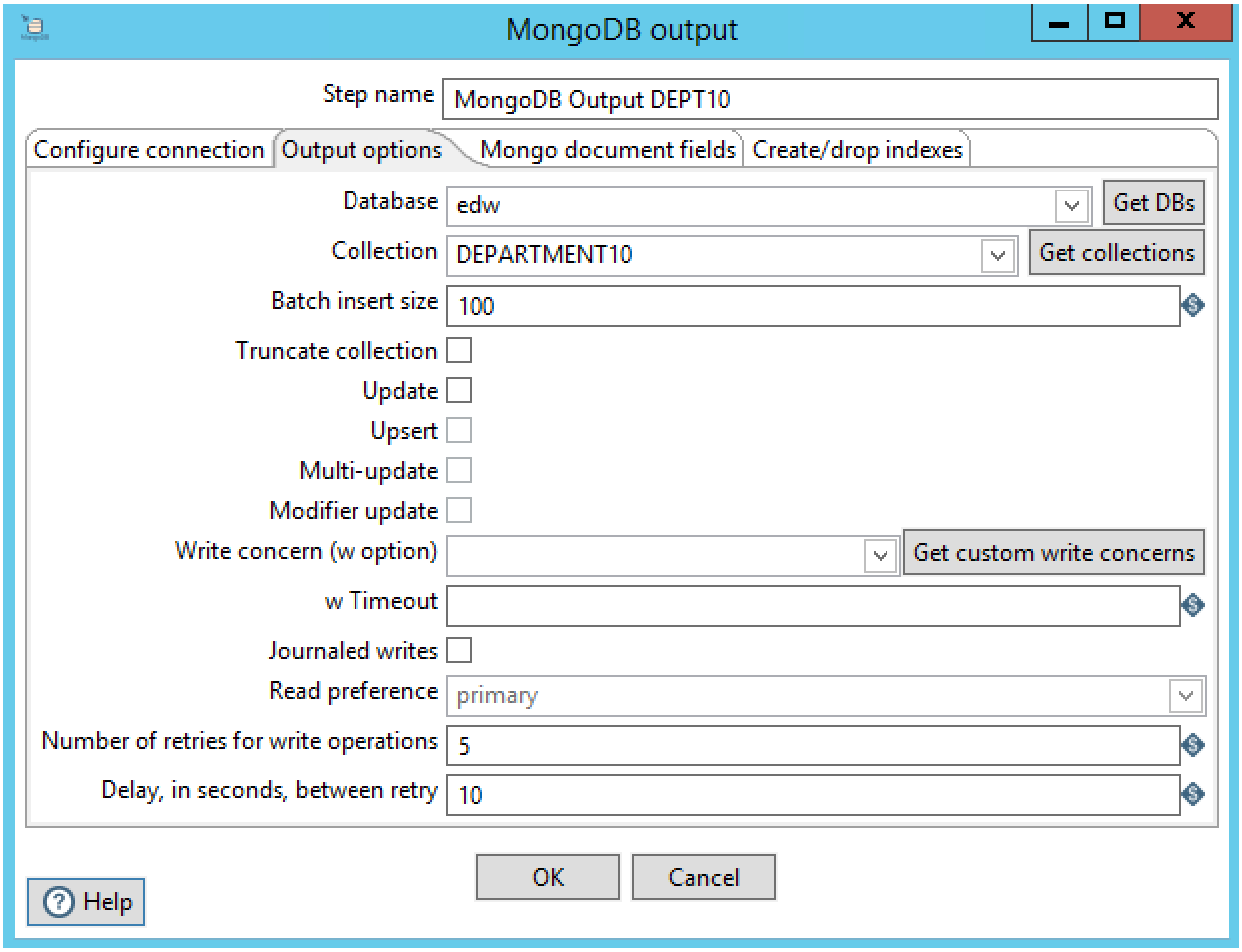
Insert data from step:

Execute for each row? □

Limit size: 0

2. To set the output options, double-click **MongoDB Output DEPT10** (Figure 3 on page 16). The **MongoDB Output** dialog box opens (Figure 5). Change values, as necessary, and click **OK**.

Figure 5



MongoDB output

Step name: MongoDB Output DEPT10

Configure connection | Output options | Mongo document fields | Create/drop indexes

Database: edw

Collection: DEPARTMENT10

Batch insert size: 100

Truncate collection □

Update □

Upsert □

Multi-update □

Modifier update □

Write concern (w option):

w Timeout:

Journaled writes □

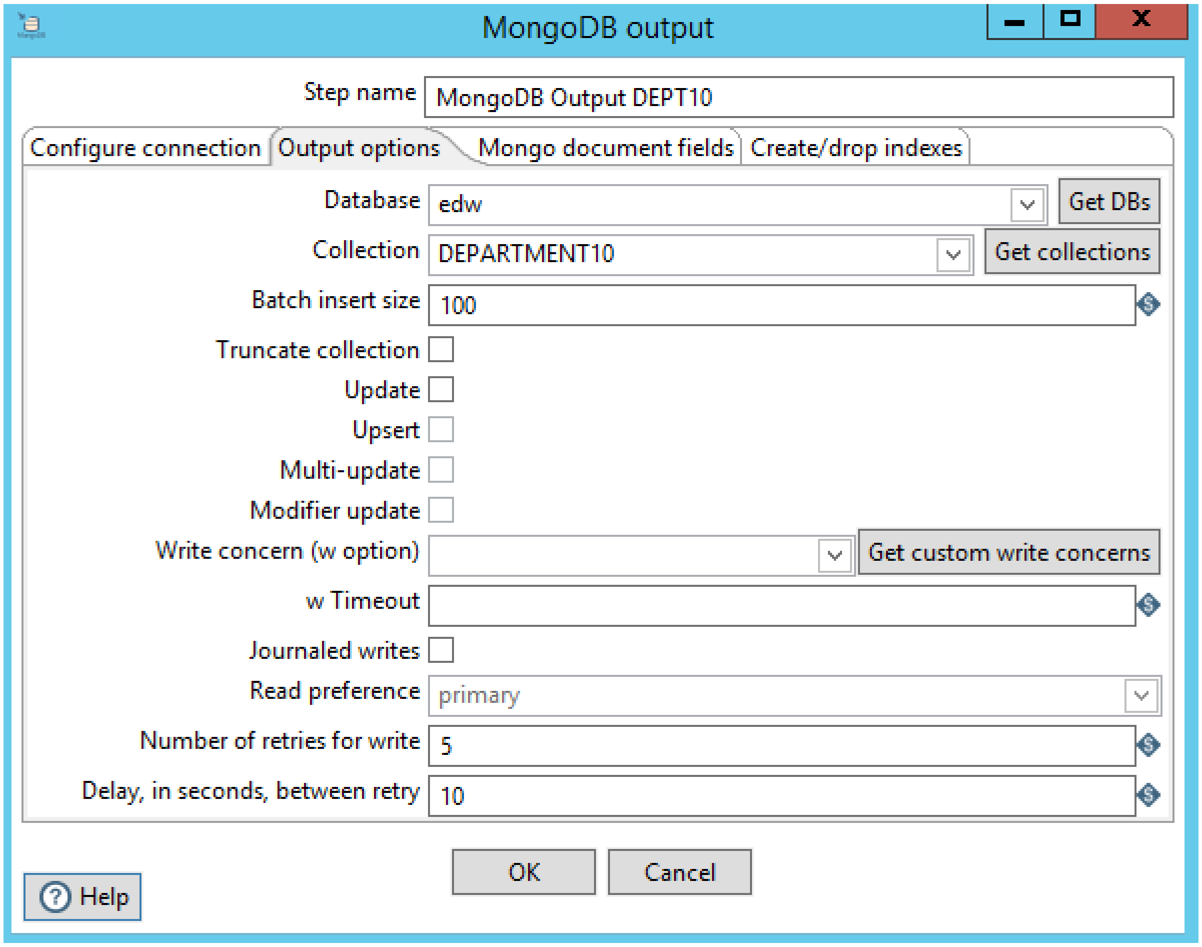
Read preference: primary

Number of retries for write operations: 5

Delay, in seconds, between retry: 10

Figure 6 shows the MongoDB document fields and the document structure that is created after clicking **OK**.

Figure 6



Step name: MongoDB Output DEPT10

Database: edw

Collection: DEPARTMENT10

Batch insert size: 100

Truncate collection ☐

Update ☐

Upsert ☐

Multi-update ☐

Modifier update ☐

Write concern (w option):

w Timeout:

Journaled writes ☐

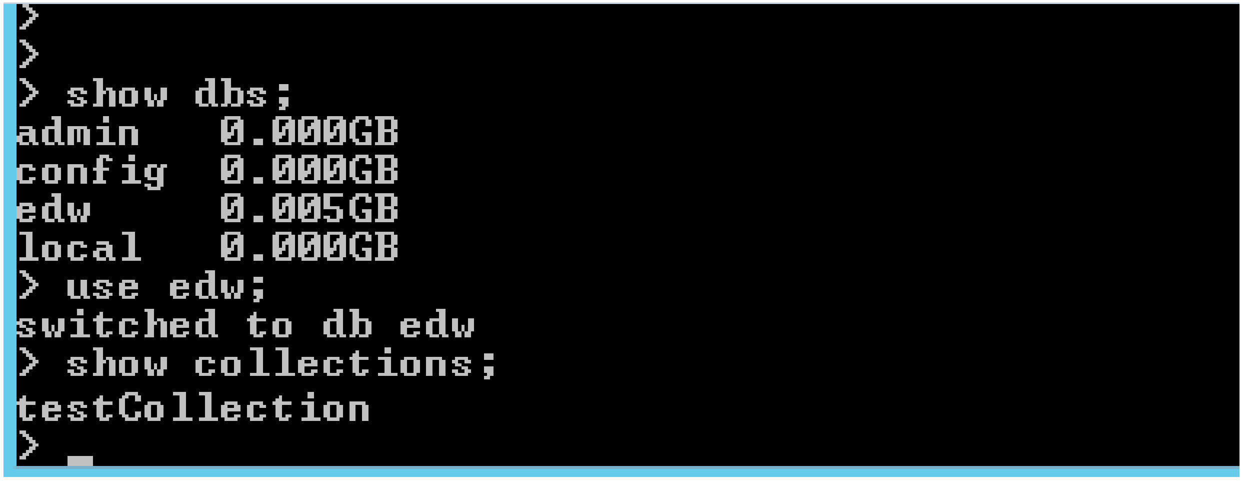
Read preference: primary

Number of retries for write: 5

Delay, in seconds, between retry: 10

Figure 7 shows the MongoDB database (collection) display before the Pentaho Data Integration workflow integration.

Figure 7

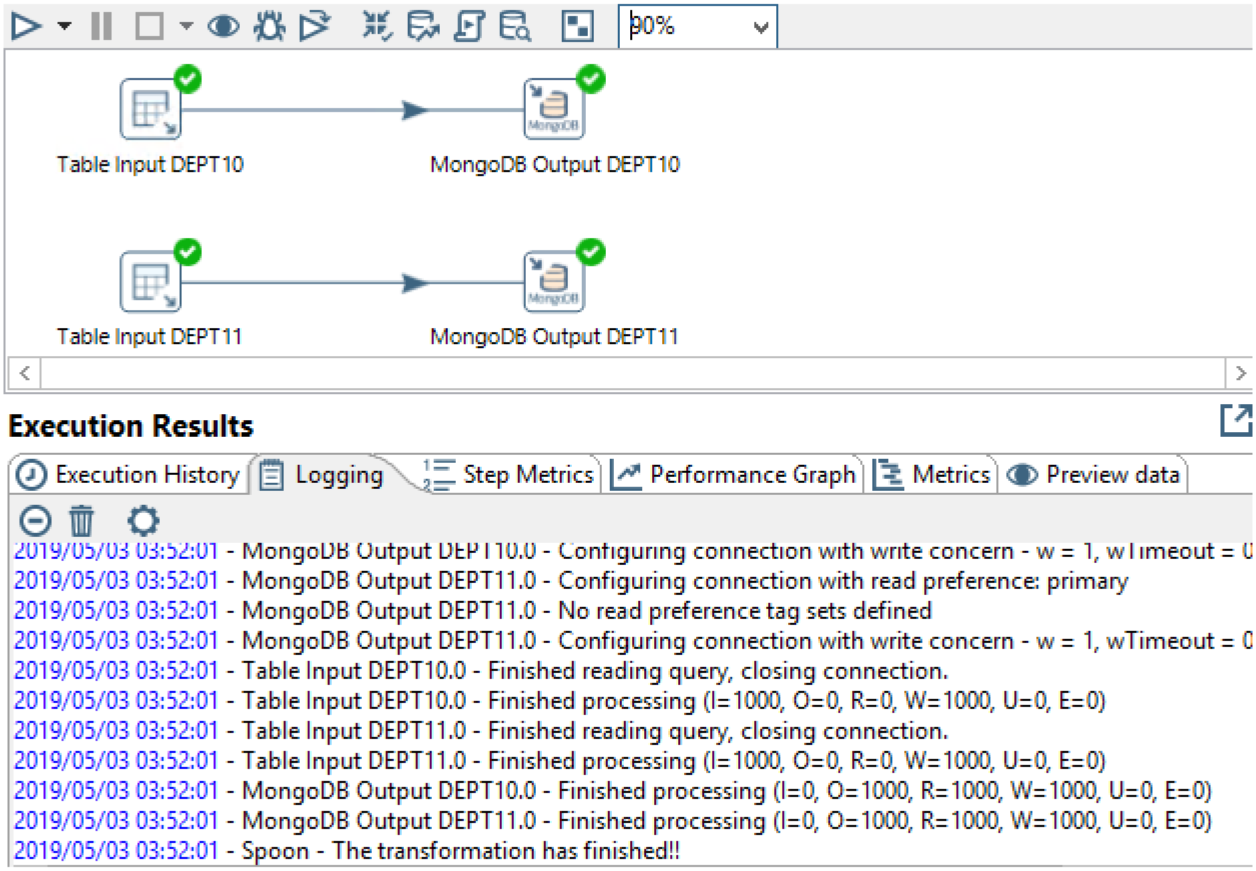


```
>
>
> show dbs;
admin   0.000GB
config  0.000GB
edw     0.005GB
local   0.000GB
> use edw;
switched to db edw
> show collections;
testCollection
> _
```

Figure 8 shows the execution results of the Pentaho Data Integration workflow. Under **Execution Results** on the **Step Metrics** tab, you can see the following results:

- The Table Input DEPT10 step shows 1000 records processed.
- The MongoDB Output DEPT10 step shows 1000 records processed.

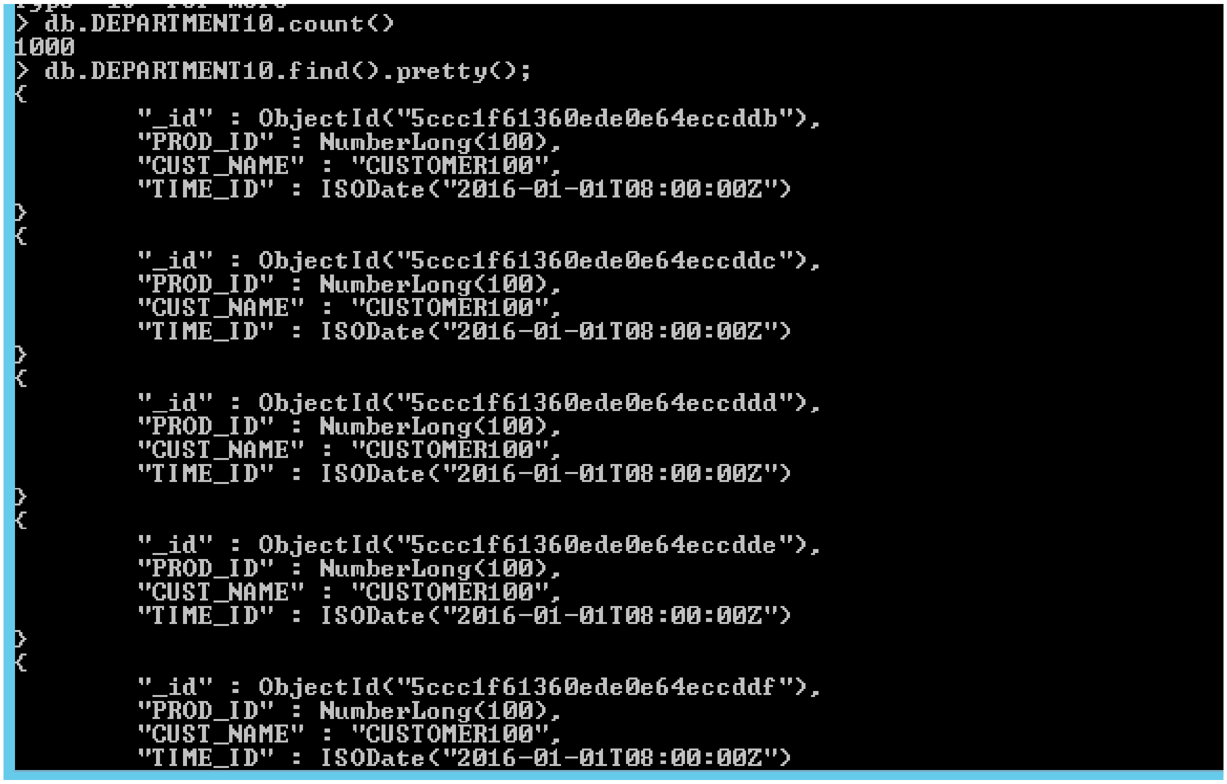
Figure 8



```
2019/05/03 03:52:01 - MongoDB Output DEPT10.0 - Configuring connection with write concern - w = 1, wTimeout = 0
2019/05/03 03:52:01 - MongoDB Output DEPT11.0 - Configuring connection with read preference: primary
2019/05/03 03:52:01 - MongoDB Output DEPT11.0 - No read preference tag sets defined
2019/05/03 03:52:01 - MongoDB Output DEPT11.0 - Configuring connection with write concern - w = 1, wTimeout = 0
2019/05/03 03:52:01 - Table Input DEPT10.0 - Finished reading query, closing connection.
2019/05/03 03:52:01 - Table Input DEPT10.0 - Finished processing (I=1000, O=0, R=0, W=1000, U=0, E=0)
2019/05/03 03:52:01 - Table Input DEPT11.0 - Finished reading query, closing connection.
2019/05/03 03:52:01 - Table Input DEPT11.0 - Finished processing (I=1000, O=0, R=0, W=1000, U=0, E=0)
2019/05/03 03:52:01 - MongoDB Output DEPT10.0 - Finished processing (I=0, O=1000, R=1000, W=1000, U=0, E=0)
2019/05/03 03:52:01 - MongoDB Output DEPT11.0 - Finished processing (I=0, O=1000, R=1000, W=1000, U=0, E=0)
2019/05/03 03:52:01 - Spoon - The transformation has finished!!
```

Figure 9 shows the MongoDB database (collection) display after executing the Pentaho Data Integration workflow.

Figure 9



```
> db.DEPARTMENT10.count()
1000
> db.DEPARTMENT10.find().pretty();
{
        "_id" : ObjectId("5ccc1f61360ede0e64eccddb"),
        "PROD_ID" : NumberLong(100),
        "CUST_NAME" : "CUSTOMER100",
        "TIME_ID" : ISODate("2016-01-01T08:00:00Z")
}
{
        "_id" : ObjectId("5ccc1f61360ede0e64eccddc"),
        "PROD_ID" : NumberLong(100),
        "CUST_NAME" : "CUSTOMER100",
        "TIME_ID" : ISODate("2016-01-01T08:00:00Z")
}
{
        "_id" : ObjectId("5ccc1f61360ede0e64eccddd"),
        "PROD_ID" : NumberLong(100),
        "CUST_NAME" : "CUSTOMER100",
        "TIME_ID" : ISODate("2016-01-01T08:00:00Z")
}
{
        "_id" : ObjectId("5ccc1f61360ede0e64eccdde"),
        "PROD_ID" : NumberLong(100),
        "CUST_NAME" : "CUSTOMER100",
        "TIME_ID" : ISODate("2016-01-01T08:00:00Z")
}
{
        "_id" : ObjectId("5ccc1f61360ede0e64eccddf"),
        "PROD_ID" : NumberLong(100),
        "CUST_NAME" : "CUSTOMER100",
        "TIME_ID" : ISODate("2016-01-01T08:00:00Z")
```

Partial Table Data Copy from Oracle to MongoDB

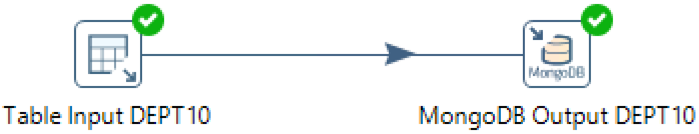
Hitachi Vantara recommends copying only the data or fields that you need in MongoDB. This results in better performance and saves storage space.

Oracle tables in a data warehouse may contain columns that are no longer useful. With the Pentaho Data Integration graphical user interface, you can explore, identify, and select specific data fields for offloading to MongoDB.

This example demonstrates how to use the user interface in Pentaho Data Integration to construct a simple workflow to copy specific columns from an Oracle table to MongoDB.

Figure 10 shows the Pentaho Data Integration workflow for a partial table data copy from Oracle to MongoDB in the user interface.

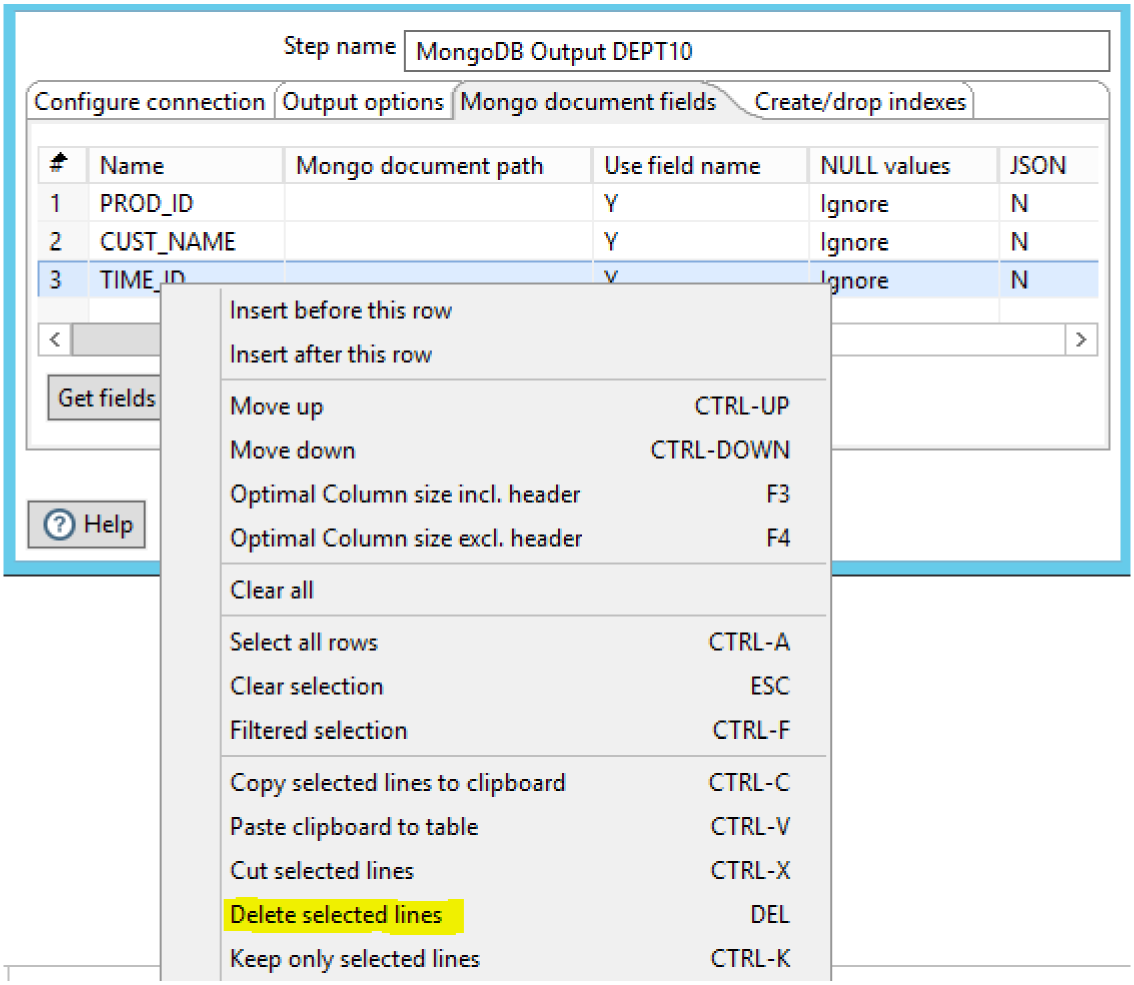
Figure 10



To create a partial table data copy workflow, do the following.

1. To exclude unnecessary columns from the export, double-click MongoDB Output DEPT10 (Figure 10 on page 20). The **MongoDB Output** dialog box opens (Figure 11). Select one or more names in the dialog box, and right-click the selected names. From the shortcut menu that opens, click **Delete selected lines**, and then click **OK**.

Figure 11



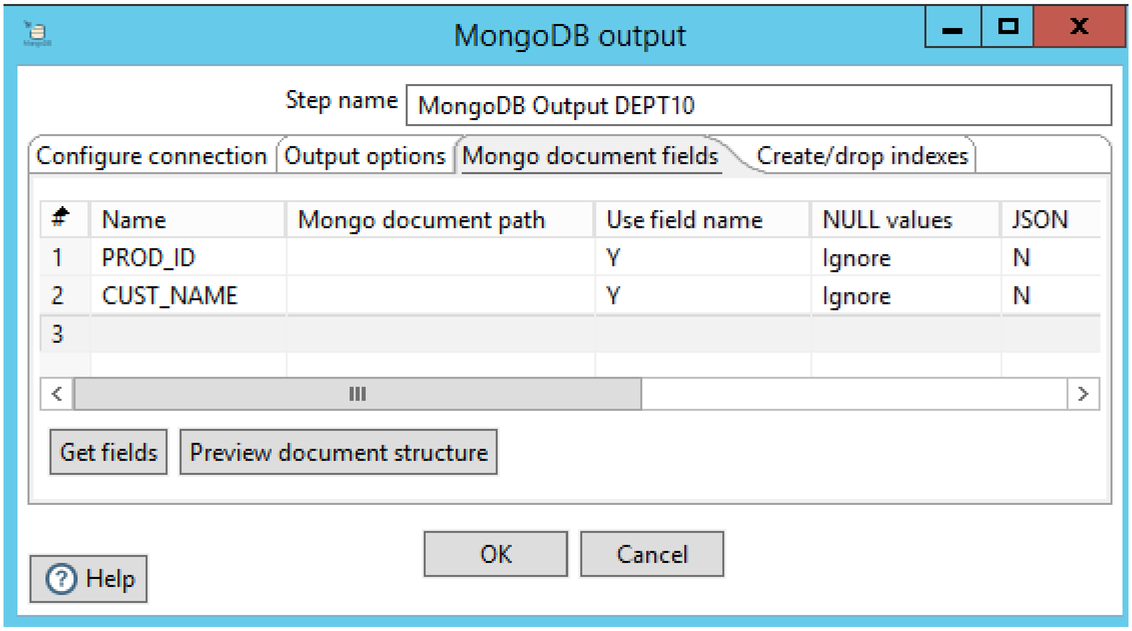
Step name: MongoDB Output DEPT10

#	Name	Mongo document path	Use field name	NULL values	JSON
1	PROD_ID		Y	Ignore	N
2	CUST_NAME		Y	Ignore	N
3	TIME_ID		Y	Ignore	N

Configure connection
Output options
Mongo document fields
Create/drop indexes
Get fields
Help
Insert before this row
Insert after this row
Move up CTRL-UP
Move down CTRL-DOWN
Optimal Column size incl. header F3
Optimal Column size excl. header F4
Clear all
Select all rows CTRL-A
Clear selection ESC
Filtered selection CTRL-F
Copy selected lines to clipboard CTRL-C
Paste clipboard to table CTRL-V
Cut selected lines CTRL-X
Delete selected lines DEL
Keep only selected lines CTRL-K

Figure 12 shows only the remaining columns to be imported from the Oracle database.

Figure 12



Step name: MongoDB Output DEPT10

#	Name	Mongo document path	Use field name	NULL values	JSON
1	PROD_ID		Y	Ignore	N
2	CUST_NAME		Y	Ignore	N
3					

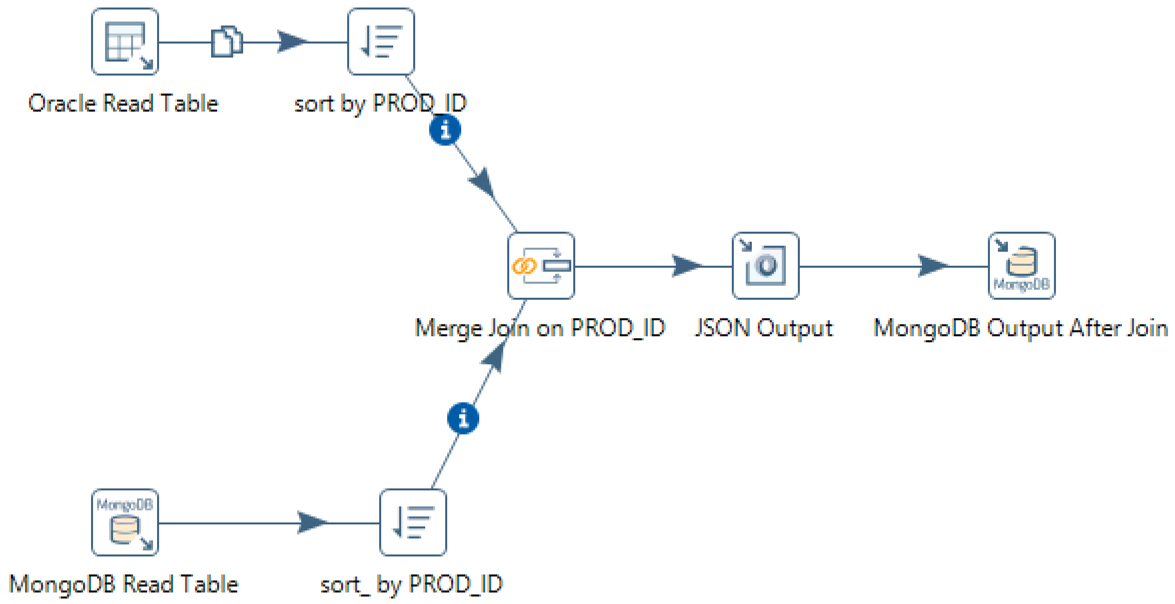
2. To return to the main user interface page, click **OK**.

Merge and Join Two Tables of Data and Copy from Oracle to MongoDB

Often you need to join two or more enterprise data warehouse tables. This example demonstrates how to use the graphical user interface in Pentaho Data Integration to join two Oracle tables and offload the data to MongoDB.

Figure 13 on page 23 shows the transformation workflow for the merged workflow in the graphical user interface of Pentaho Data Integration.

Figure 13



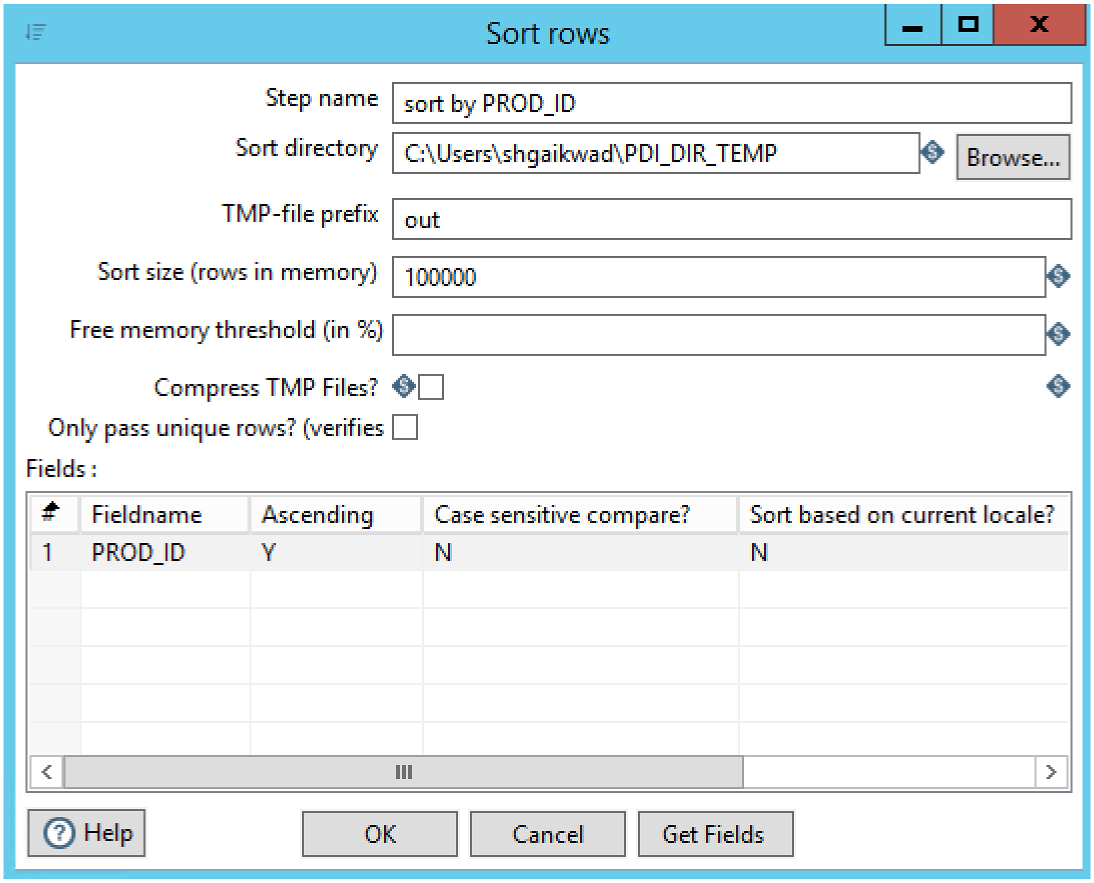
As part of the Pentaho workflow, sorting rows in large tables can be time consuming. Hitachi Vantara recommends sorting all the rows in server memory instead of using a memory-plus-disk approach.

On the **Sort row** dialog box, make the following settings:

- Use the **Sort size (rows in memory)** text box to control how many rows are sorted in server memory.
- Use the **Free memory threshold (in%)** text box to help avoid filling all available memory in server memory. Make sure to allocate enough RAM to Pentaho Data Integration on the server when you need to do large sorting tasks.

Figure 14 shows controlling the cache size in the **Sort rows** dialog box from the graphical user interface in Pentaho Data Integration.

Figure 14



Step name: sort by PROD_ID

Sort directory: C:\Users\shgaikwad\PDI_DIR_TEMP

TMP-file prefix: out

Sort size (rows in memory): 100000

Free memory threshold (in %):

Compress TMP Files? □

Only pass unique rows? (verifies □

Fields :

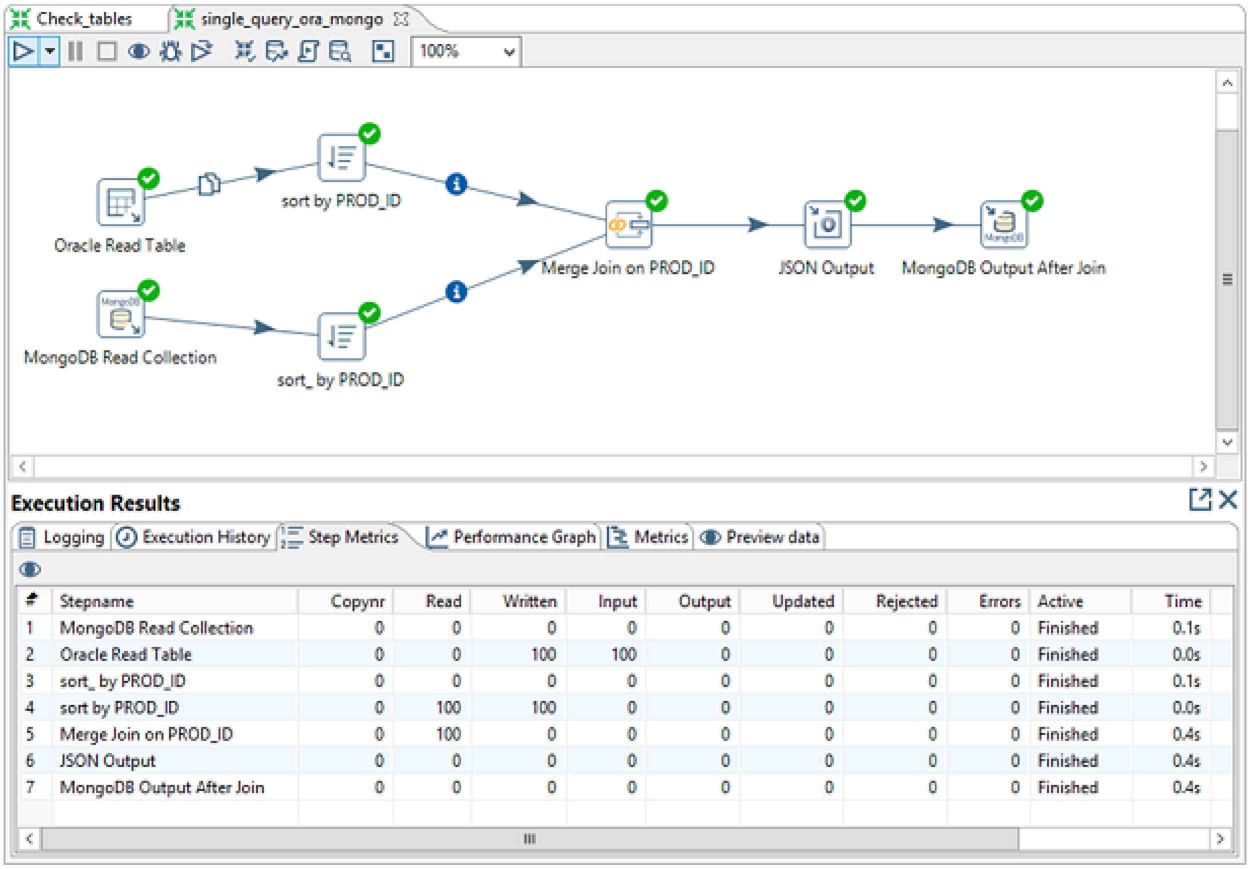
#	Fieldname	Ascending	Case sensitive compare?	Sort based on current locale?
1	PROD_ID	Y	N	N

Sorting on the database is faster often than sorting externally, especially if there is an index on the sort field or fields. You can use this as another option to improve performance.

More Oracle tables can be joined one at time with same Pentaho Data Integration sorting and join steps. You can also use **Execute SQL Script** in Pentaho for another option to join multiple tables. [This example in the Pentaho Community Forums] shows how to do this from Pentaho Data Integration.

Figure 15 shows the transformation workflow and the execution results for the merged workflow in the graphical user interface of Pentaho Data Integration. This example shows one data source is the Oracle database table and other is the MongoDB collection. Pentaho Data Integration receives a single query, sends the query to both databases, and joins the received results. The joined result is saved to MongoDB for future analytics.

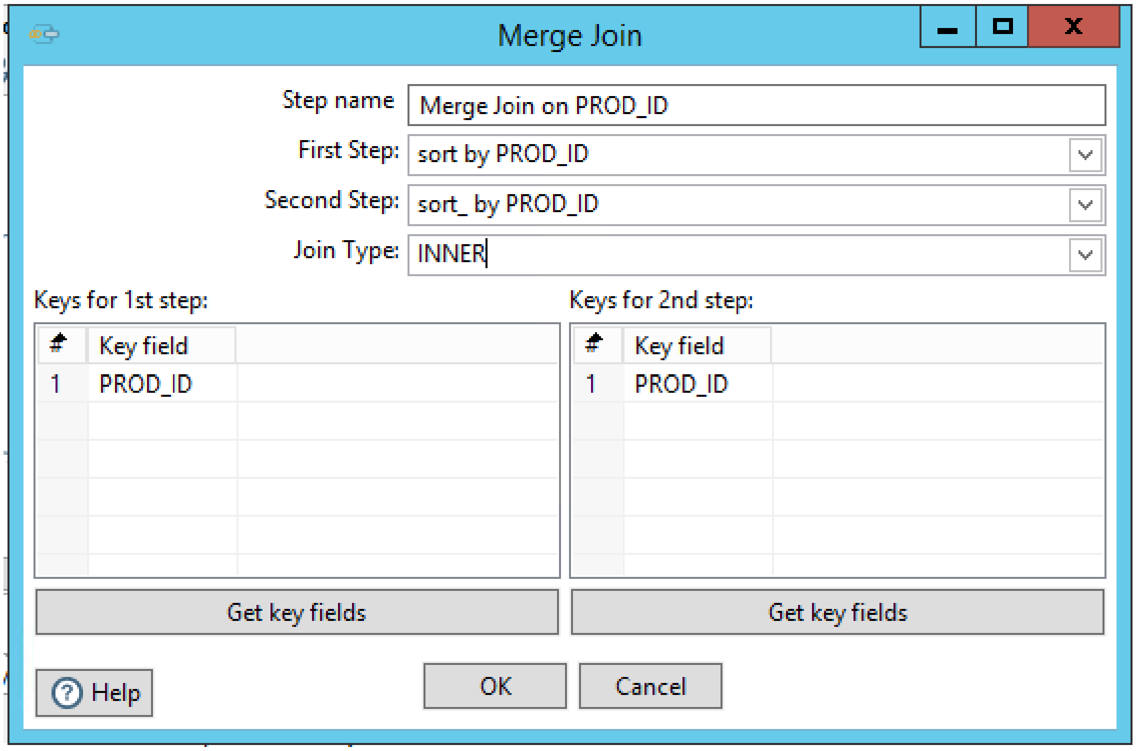
Figure 15



#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time
1	MongoDB Read Collection	0	0	0	0	0	0	0	0	Finished	0.1s
2	Oracle Read Table	0	0	100	100	0	0	0	0	Finished	0.0s
3	sort_ by PROD_ID	0	0	0	0	0	0	0	0	Finished	0.1s
4	sort by PROD_ID	0	100	100	0	0	0	0	0	Finished	0.0s
5	Merge Join on PROD_ID	0	100	0	0	0	0	0	0	Finished	0.4s
6	JSON Output	0	0	0	0	0	0	0	0	Finished	0.4s
7	MongoDB Output After Join	0	0	0	0	0	0	0	0	Finished	0.4s

Figure 16 shows the merge join details of the example.

Figure 16



Merge Join

Step name: Merge Join on PROD_ID

First Step: sort by PROD_ID

Second Step: sort_ by PROD_ID

Join Type: INNER

Keys for 1st step:

#	Key field
1	PROD_ID

Keys for 2nd step:

#	Key field
1	PROD_ID

Performance Measurement

Table 8 shows script execution results, measuring the performance of the offload operation. Offload performance depends on numerous factors, including database tuning, operating system configuration, network, and storage configuration. These sample results are general guidelines only and might not be realized in your implementation.

TABLE 8. PERFORMANCE MEASUREMENT EXAMPLE

Number of Records Offloaded	1 Million	3 Million	8 Million
Sequential offloading time in minutes [Note 1]	2.03	6.01	15.54

[Note 1] Offloading in this case was the time it took to copy the data, rather than copying and erasing from the source. Copying and erasing is still possible through configuration changes following your requirements.

Engineering Validation

This summarizes the key observations from the test results for Hitachi Solution for Databases in an enterprise data warehouse offload package for Oracle Database with Hitachi Virtual Hitachi Advanced Server DS120, Advanced Server DS220, Pentaho 8.3, and MongoDB 4.2.

When evaluating this Oracle Enterprise Data Warehouse (EDW) solution, the laboratory environment used the following:

- One Hitachi Unified Compute Platform CI for the Oracle Database environment
- One Hitachi Advanced Server DS220
- One Hitachi Virtual Storage Platform G600
- Two Brocade G620 SAN switches

Using this same test environment is not a requirement to deploy this solution.

Test Methodology

The source data was preloaded into a sample database schema into Oracle database.

After preloading the data, a few example Pentaho Data Integration workflows were developed for offloading data to a MongoDB database.

Once data was loaded into the MongoDB database, verification was done to make sure the data was offloaded correctly.

The example workflows found in Enterprise Data Offload Workflow were used to validate this environment.

Testing involved following this procedure for each example workflow:

1. Verify the following for each Oracle table before running the enterprise data offload workflow:
   - Number of rows
   - Number of columns
   - Data types
2. Run the enterprise data offload workflow.
3. Verify the following for the MongoDB collection after running the enterprise data offload workflow to see if the numbers matched those in the Oracle table:
   - Number of documents (same as number of rows)
   - Number of fields (same as number of columns)
   - Data types

The Python toolkit, as mentioned in Enterprise Data Offload Workflow, can be used to prepare Pentaho Data Integration workflows to very quickly copy Oracle data to MongoDB. Apart from MongoDB version 4.2, the toolkit also supports other MongoDB versions, such as 3.2 and 3.6.

Test Results

After running each enterprise data offload workflow example, the test results showed the same number of documents (rows), fields (columns), and data types in the Oracle database as the MongoDB database.

These results show that you can use Pentaho Data Integration to move data from an Oracle host to MongoDB hosts to relieve the workload on your Oracle host. This provides a cost-effective solution to expanding capacity to relieve server utilization pressures.

For More Information

Hitachi Vantara Global Services offers experienced storage consultants, proven methodologies and a comprehensive services portfolio to assist you in implementing Hitachi products and solutions in your environment. For more information, see the Services website.

Demonstrations and other resources are available for many Hitachi products. To schedule a live demonstration, contact a sales representative or partner. To view on-line informational resources, see the Resources website.

Hitachi Academy is your education destination to acquire valuable knowledge and skills on Hitachi products and solutions. Our Hitachi Certified Professional program establishes your credibility and increases your value in the IT marketplace. For more information, see the Hitachi Vantana Training and Certification website.

For more information about Hitachi products and services, contact your sales representative, partner, or visit the Hitachi Vantara website.

Hitachi Vantara

Corporate Headquarters
2845 Lafayette Street
Santa Clara, CA 95054 USA
www.HitachiVantara.com | community.HitachiVantara.com

Regional Contact Information
USA: 1-800-446-0744
Global: 1-858-547-4526
HitachiVantara.com/contact

Notice: This document is for informational purposes only, and does not set forth any warranty, expressed or implied, concerning any equipment or service offered or to be offered by Hitachi Vantara.

MK-SL-177-00. November 2019.