



Informatica®
10.4.1

Data Quality Performance Tuning Guide

This software and documentation are provided only under a separate license agreement containing restrictions on use and disclosure. No part of this document may be reproduced or transmitted in any form, by any means (electronic, photocopying, recording or otherwise) without prior consent of Informatica LLC.

Informatica and the Informatica logo are trademarks or registered trademarks of Informatica LLC in the United States and many jurisdictions throughout the world. A current list of Informatica trademarks is available on the web at <https://www.informatica.com/trademarks.html>. Other company and product names may be trade names or trademarks of their respective owners.

Subject to your opt-out rights, the software will automatically transmit to Informatica in the USA information about the computing and network environment in which the Software is deployed and the data usage and system statistics of the deployment. This transmission is deemed part of the Services under the Informatica privacy policy and Informatica will use and otherwise process this information in accordance with the Informatica privacy policy available at <https://www.informatica.com/in/privacy-policy.html>. You may disable usage collection in Administrator tool.

U.S. GOVERNMENT RIGHTS Programs, software, databases, and related documentation and technical data delivered to U.S. Government customers are "commercial computer software" or "commercial technical data" pursuant to the applicable Federal Acquisition Regulation and agency-specific supplemental regulations. As such, the use, duplication, disclosure, modification, and adaptation is subject to the restrictions and license terms set forth in the applicable Government contract, and, to the extent applicable by the terms of the Government contract, the additional rights set forth in FAR 52.227-19, Commercial Computer Software License.

Portions of this software and/or documentation are subject to copyright held by third parties. Required third party notices are included with the product.

See patents at <https://www.informatica.com/legal/patents.html>.

The information in this documentation is subject to change without notice. If you find any problems in this documentation, report them to us at infa_documentation@informatica.com.

Informatica products are warranted according to the terms and conditions of the agreements under which they are provided. INFORMATICA PROVIDES THE INFORMATION IN THIS DOCUMENT "AS IS" WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

Publication Date: 2022-11-10

Table of Contents

Preface	5
Informatica Resources.	5
Informatica Network.	5
Informatica Knowledge Base.	5
Informatica Documentation.	5
Informatica Product Availability Matrices.	6
Informatica Velocity.	6
Informatica Marketplace.	6
Informatica Global Customer Support.	6
 Chapter 1: General Sizing Guidelines.....	 7
Overview.	7
Basic On-Disk Installation Size.	7
General runtime memory size.	8
Address Validation Reference Data.	8
Capacity Planning.	9
 Chapter 2: Data Quality Transformations.....	 11
General Performance Guidelines.	11
Address Validator.	11
Performance Guidelines for Address Validation.	12
Association.	13
Consolidation.	13
Key Generator.	14
Match.	14
 Chapter 3: Data Quality Component Configuration.....	 17
Overview.	17
Human Tasks.	17
Workflow Orchestration Service Connection Pooling.	17
Data Integration Service Java Heap Size.	18
Human Task Load Balancing.	18
Probabilistic Models.	18
Design-Time Considerations.	18
Model Creation.	19
Classifier Models.	19
Design-Time Considerations.	19
Model Creation.	19
Hadoop.	20
Web Services.	20

Multithreading.	21
-------------------------	----

Preface

Refer to the *Informatica® Data Quality Performance Tuning Guide* to learn how to configure your data quality transformations to optimize run-time performance and to learn about the memory requirements that apply to Informatica reference data.

Informatica Resources

Informatica provides you with a range of product resources through the Informatica Network and other online portals. Use the resources to get the most from your Informatica products and solutions and to learn from other Informatica users and subject matter experts.

Informatica Network

The Informatica Network is the gateway to many resources, including the Informatica Knowledge Base and Informatica Global Customer Support. To enter the Informatica Network, visit <https://network.informatica.com>.

As an Informatica Network member, you have the following options:

- Search the Knowledge Base for product resources.
- View product availability information.
- Create and review your support cases.
- Find your local Informatica User Group Network and collaborate with your peers.

Informatica Knowledge Base

Use the Informatica Knowledge Base to find product resources such as how-to articles, best practices, video tutorials, and answers to frequently asked questions.

To search the Knowledge Base, visit <https://search.informatica.com>. If you have questions, comments, or ideas about the Knowledge Base, contact the Informatica Knowledge Base team at KB_Feedback@informatica.com.

Informatica Documentation

Use the Informatica Documentation Portal to explore an extensive library of documentation for current and recent product releases. To explore the Documentation Portal, visit <https://docs.informatica.com>.

If you have questions, comments, or ideas about the product documentation, contact the Informatica Documentation team at infa_documentation@informatica.com.

Informatica Product Availability Matrices

Product Availability Matrices (PAMs) indicate the versions of the operating systems, databases, and types of data sources and targets that a product release supports. You can browse the Informatica PAMs at <https://network.informatica.com/community/informatica-network/product-availability-matrices>.

Informatica Velocity

Informatica Velocity is a collection of tips and best practices developed by Informatica Professional Services and based on real-world experiences from hundreds of data management projects. Informatica Velocity represents the collective knowledge of Informatica consultants who work with organizations around the world to plan, develop, deploy, and maintain successful data management solutions.

You can find Informatica Velocity resources at <http://velocity.informatica.com>. If you have questions, comments, or ideas about Informatica Velocity, contact Informatica Professional Services at ips@informatica.com.

Informatica Marketplace

The Informatica Marketplace is a forum where you can find solutions that extend and enhance your Informatica implementations. Leverage any of the hundreds of solutions from Informatica developers and partners on the Marketplace to improve your productivity and speed up time to implementation on your projects. You can find the Informatica Marketplace at <https://marketplace.informatica.com>.

Informatica Global Customer Support

You can contact a Global Support Center by telephone or through the Informatica Network.

To find your local Informatica Global Customer Support telephone number, visit the Informatica website at the following link:

<https://www.informatica.com/services-and-training/customer-success-services/contact-us.html>.

To find online support resources on the Informatica Network, visit <https://network.informatica.com> and select the eSupport option.

CHAPTER 1

General Sizing Guidelines

This chapter includes the following topics:

- [Overview, 7](#)
- [Basic On-Disk Installation Size, 7](#)
- [General runtime memory size, 8](#)
- [Capacity Planning, 9](#)

Overview

You can verify and enhance the performance of your Informatica Data Quality and Data Engineering Quality installations by monitoring and updating key parameters in your Informatica domain and applications. Factors affecting performance include the memory available to the applications and the properties that you select on the transformations and mappings that you configure.

For additional information on Data Quality system requirements, consult the [Installation Guide for PowerCenter and Data Quality](#) on the Informatica documentation portal.

Basic On-Disk Installation Size

When you evaluate your resource needs, consider the number of concurrent mappings you will run, the types of transformation used in each mapping, and the size of the source data sets.

The following table provides a guide to the reference data footprint in a standard installation

Reference Data Type	Size on Disk
Identity reference data files	600 MB
Address reference data files for batch/interactive validation	9 GB
Reference table data	3 GB

The installation of additional reference will affect these figures.

The following table describes the disk space necessary for additional address reference data files:

Reference Data Type	Size on Disk
Geocoding	11 GB
Fast completion	7 GB
Consumer segmentation	2 GB

General runtime memory size

Before you install Informatica Data Quality, consider the amount of memory you will need. The following table describes a standard installation that is not running disk- or memory-intensive mappings or mappings that load reference data.

Service Name	Virtual Set	Working Set
Administrator Tool	1229 KB	506 KB
Model Repository Service	2041 KB	912 KB
Data Integration Service	1507 KB	474 KB
Analyst Tool	1422 KB	570 KB
Content Management System	977 KB	315 KB

The Virtual Set is the total virtual memory used, and the Working Set is the physically resident memory used.

Address Validation Reference Data

Use the address reference data file sizes as a guide to memory usage.

The Content Management Service dictates how the address reference data files are loaded. You can view the Content Management Service settings in the Administrator tool. The Data Integration Service applies the Content Management Service settings when it loads the address reference data. The Data Integration Service loads the data in the same way for all users.

The average size in memory of each loaded element is approximately the same as the disk footprint. For example, if a user runs a mapping that uses a 533 MB address reference data file in batch or interactive mode, the process memory size will grow by approximately 533 MB.

The memory remains in use while any address validation mapping runs. The Data Integration Service unloads the address reference data and releases the memory when the mapping finishes.

Capacity Planning

You can use the metrics that follow to estimate the capacity of your enterprise. The performance metrics are measured on a system running Informatica 10.4.1 with the following specifications:

System Element	Informatica Data Quality	Data Engineering Quality
Domain	16 Core (4CPUs) 62 GB RAM	16 Core (4CPUs) 62 GB RAM
DIS Heap size	4 GB	4 GB
Cluster size	Not Applicable	6 Node CDH Cluster 76 GB RAM

Capacity Planning in Informatica Data Quality

The following table shows the performance results for a range of mappings in Informatica Data Quality:

Mapping	Number of Rows	Time (milliseconds)	Rows per Second
Address Validator transformation - Great Britain data	999999	2365653	422
Address Validator transformation - United States data	3000000	2999056	1000
Match transformation - field matching	3000000	558385	5372
Match transformation - identity matching	250000	274424	910
Match transformation - identity matching - load data to persistent index	1000000	681343	1467
Match transformation - identity matching - incremental update to persistent index.	10000	123764	80
Parse transformation configured with probabilistic models	200000	737750	271
Parse transformation configured with probabilistic models - 8 partitions enabled	200000	269248	742
Standardizer transformation and Merge transformation	335999987	577715	581601
Standardizer transformation and Merge transformation - 8 partitions enabled	335999987	227822	1474835

Capacity Planning in Data Engineering Quality

The following table shows the performance results for a range of mappings in Data Engineering Quality:

Mapping	Number of Rows	Time on Spark(milliseconds)	Time on Blaze(milliseconds)	Rows per Second (Spark)	Rows per Second (Blaze)
Address Validator transformation - Great Britain data	999999	116229	595951	8603	1677
Address Validator transformation - United States data	3000000	167421	1432382	17918	2094
Case Converter transformation configured with reference tables	3000000	163703	326450	18325	9189
Key Generator transformation	3000000	84810	10822	35373	277213
Match transformation - field matching	3000000	165392	613653	18138	4888
Match transformation - identity matching	250000	1139996	2274785	219	109
Merge transformation	335999987	485996	327536	691363	1025841
Parse transformation configured with probabilistic models	200000	446415	416932	448	479

CHAPTER 2

Data Quality Transformations

This chapter includes the following topics:

- [General Performance Guidelines, 11](#)
- [Address Validator, 11](#)
- [Association, 13](#)
- [Consolidation, 13](#)
- [Key Generator, 14](#)
- [Match, 14](#)

General Performance Guidelines

You can configure your data quality transformations in several ways to improve performance.

Consider the following general guidelines for transformation performance:

- Set the maximum length on string ports accurately. Do not set values that far exceed the physical size of the data that the ports will carry.
- Configure your input ports to be the same type as the corresponding output ports. For example, cast numbers to numbers and strings to strings. Casting from one type to another when not necessary will have a performance impact.
- Ensure that the Tracing Level in all transformations is set to Normal, as performance is degraded when set to a more verbose options.

Address Validator

Several configuration options affect the performance of the Address Validator transformation. You can review and update the configuration properties on the Content Management Service.

Preloading Method

When the preloading method is set to MAP, Data Quality does not load the address reference data for every process that uses it. Instead, the reference data is shared across processes. This is significant when more than one Data Integration Service process or job is set to run out-of-process. If the reference data is preloaded for another process, it will not be loaded again.

Memory Usage

The Memory Usage option controls the amount of memory that the Data Integration Service can use to preload address reference data. If the amount of memory is insufficient to preload the required data, the Data Integration Service will attempt to partially preload the data. If the amount of available memory is too low, the service will not preload any data.

Max Address Object Count

The Max Address Object Count option sets the maximum number of address validation instances that can run concurrently in a Data Integration Service process. If you run a mapping that includes an Address Validator transformation and all instances are in use, the Data Integration Service generates a server error. The Data Integration Service also returns this error if the number of Address Validator transformations in the mapping breaches the maximum number of concurrent instances.

Max Thread Count

The Max Thread Count option determines the number of separate threads that Address Validator instances can use. As best practice, set the option value to one less than the total number of threads or cores on the Data Integration Service machine. Setting a larger value will consume memory with no performance benefit. Performance is adversely impacted if the number of address verification instances exceeds the number of available threads.

Cache Size

The Cache Size value affects the PARTIAL and NONE preloading options and can yield a small improvement in performance. Increasing the cache size can improve the loading of address data, particularly below the Locality level.

Performance Guidelines for Address Validation

Consider the following rules and guidelines when you configure your system for address validation:

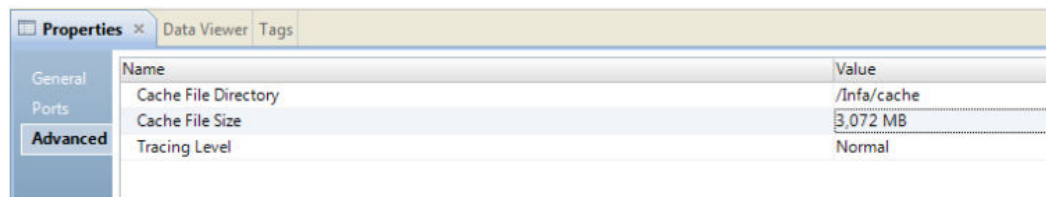
- You can store the reference data on a fast hard disk, solid-state disk, or even a flash disk (high-speed USB stick).
- Where possible, install sufficient memory to allow all databases to fully pre-load into memory.
- Preload at least the databases of frequently used countries. At a minimum, the available memory should equal the aggregate size of the most often-used country databases plus 256 MB.
- If you will use reference data from all countries simultaneously, add memory to cover the size of the databases.
- Use a 64-bit environment to preload more than 3 GB of reference data.
- Do not set a country code as a No Preload value. Enter a No Preload value of ALL to avoid using a country code.
- Minimize the access latency (average access time).
- If you use a solid-state disk, do not preload the databases. Set a LARGE cache size in the Content Management Service instead.
- Do not use the same drive to store address reference data and source or target files.
- When enough memory is available, processor speed directly determines the speed of address processing.
- Try to sort your address records by country or postcode prior to processing. Validation also benefits from internal and operating system caches for sorted addresses as opposed to addresses in random order.
- The Max Thread Count value must be greater than or equal to the number of partitions.

- Configure the Execution Instances property on the Address Validator transformation in conjunction with the Max Thread Count value.

Association

If you run the Association transformation on a large data set, the transformation may not be able to store all associated records in memory, and some records will be written to disk. The Cache File Size property on the transformation specifies the amount of memory available.

The following image shows the property:



A cache size value below 65536 represents megabytes, and any higher value represents bytes.

The Cache File Directory identifies a storage area for the temporary files that the association operations create. Configure the cache directory on the smallest, fastest disk for performance improvements.

B-Tree Considerations

The Association transformation makes extensive use of B-tree file-based storage. Each column that the transformation reads has its own B-tree, and a general B-tree is used to store all input data rows. The Informatica B-tree is space-efficient but not compressed.

Use the following formulas to determine the needs of the transformation:

- Association transformation column size:
Total volume of data for each column + 20 bytes for each input row
- On-disk runtime cost of the general storage cache:
Size of input data set + 10 bytes for each row
- Maximum internal memory map for association IDs and data rows:
Number of rows + 20 bytes

Note: You cannot partition the Association transformation.

Consolidation

The Consolidation transformation uses standard Informatica sorting techniques. By default, the techniques give the transformation as much memory as possible without affecting system performance.

You can set a limit on the amount of main memory that the transformation uses to sort data. This increases on-disk temporary memory use, as the transformation must store all data rows.

Note: You cannot partition a Consolidation transformation.

Key Generator

The Key Generator transformation can create a set of unique identifiers for the rows in a data set. Use the transformation to create sequence ID values for a Match transformation.

The Key Generator transformation includes an option to sort the output data. To maximize performance, do not check this option. To sort the data, pass the output to a Sorter transformation and configure the cache settings on the Sorter.

Match

To optimize performance in a Match transformation, you must understand the concepts that underpin match analysis.

Single-Source field matching

Single-source field match analysis compares data from every record in a data set with every other record. The analysis generates a numerical score for every pair of records that it compares. To reduce processing time, the transformation uses one or more Key fields to organize the input records into groups prior to match analysis. You select the Key fields. The number of record pairs created depends on the number of records within a group.

The number can be calculated by the following formula:

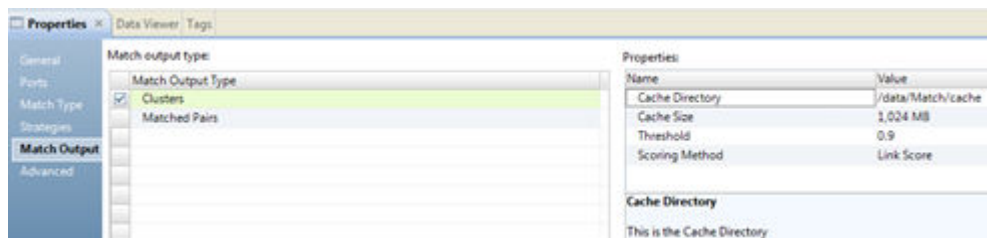
$$\frac{(n^2 - n)}{2} \text{ where } n \text{ is the number of records in the group.}$$

Group size has a significant impact on performance. For example, applying the formula above to a group of 2,000 records will produce 1,999,000 matches. Applying the formula to a group of 5,000 records will produce 12,497,500 matches, or over six times the amount.

For optimal performance, groups of over 10,000 are not recommended. Group sizes should be meaningful, so that you do not miss possible matches, but they should not be too large.

If you perform matching on a large data set, the Match transformation may not be able to store all comparison pairs in memory, and some pairs will be written to disk. The Cache Size property on the transformation determines the amount of memory available.

The following image shows the property:



A cache size value below 65536 is measured in megabytes, and any higher value is measured in bytes.

The Cache Directory property identifies a storage area for the temporary files that match analysis creates. Configure the cache directory on the smallest, fastest disk for performance improvements.

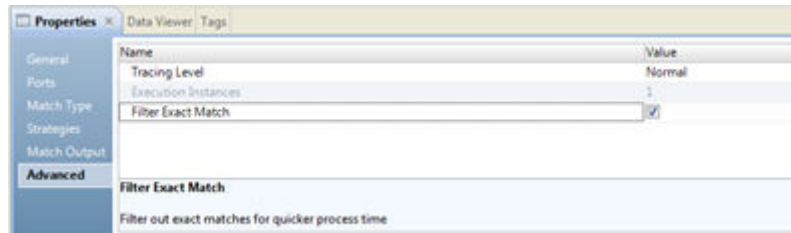
Where possible, do not use pass-through ports on the Match transformation, especially in large data sets. The pass-through ports take up valuable memory or disk space. To reunite the ports with the matched records, you can use a Joiner transformation that reads the sequence ID values.

The Match transformation can generate Link Score and Driver Score values that represent the degrees of similarity between different pairs of records in a cluster of matching records.

For optimum performance, choose Link Scores and not Driver Scores. Choosing Driver Scores will greatly decrease the performance of your match mapping, as Driver Scores write more information to disk.

Selecting the Filter Exact Match property significantly improves match performance if the data contains a significant number of exactly matched pairs. Otherwise the option has a negligible performance impact.

The following image shows the Filter Exact Match property:



Dual-Source Field Matching

Many of the principles of single-source matching also apply in dual-source matching. However, in dual-source matching, the Match transformation compares each record in one data set with every record in the other data set.

The following formula calculates the number of pairs:

$n \times m$, where n is the number of records in group 1 in data set 1 and m is number of records in group 1 in data set 2.

For example, if data set 1 includes a group with 3,000 rows and the same group exists in data set 2 with 2,000 rows, match analysis will generate 6,000,000 record pairs.

Identity Matching

The use of groups in identity matching is optional but recommended. As is the case in field matching, very large group sizes will result in considerably slower performance.

To significantly improve identity matching performance, increase the number of execution instances on the transformation. When you increase the number of execution instances, the Data Integration Service splits the workload over multiple threads. The availability of execution instances depends on the number of processor cores on the Data Integration Service machine.

The performance improvement will not be linear. The complete matching process cannot be split over multiple threads. Part of the process must be completed in a single thread.

Note: For optimal performance with identity matching, set your execution instances to the number of processor cores minus 1.

Disk Space Required

Consider the following factors for disk space sizing in field and identity matching:

Field matching

The following formula is a guide to the quantity of disk space in MB required to run field matching on a data set, generating only the link score:

$$d * n * 0.0000025$$

where d = the sum of the Match transformation input port precisions, n = the number of records, and 0.0000025 = the memory required per character.

If the mapping has dual sources, n in the above formula represents the total of the two sources.

The result above will double when the driver score is required.

Identity matching

The following formula is a guide to the quantity of disk space in MB required to run identity matching on a data set, generating only the link score:

$$d * n * 0.000005$$

where d = the sum of the Match transformation input port precisions, n = the number of records, and 0.000005 = the memory required per character.

If the mapping has dual sources, n in the above formula represents the total of the two sources.

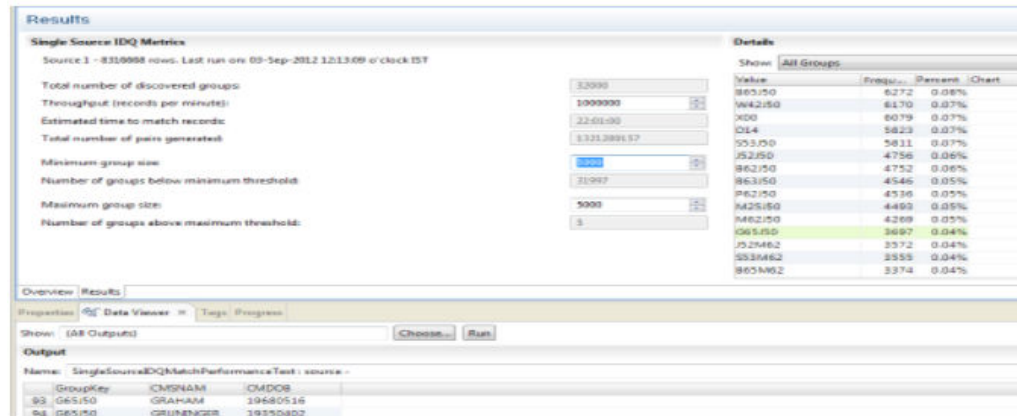
Match Performance Analysis

The Match Performance Analysis feature, available as a right-click option on the Match transformation, is valuable in estimating how long a mapping with a Match transformation may take to complete.

The report represents a profile of the data and including a table that describes the composition of the groups.

Note: The Developer tool shows the first 16,000 groups. To see the full makeup of the data, export the report to a file.

The following image shows a match performance analysis report:



Modify the minimum and maximum group sizes to evaluate the likely effect of different group sizes on the mapping performance.

CHAPTER 3

Data Quality Component Configuration

This chapter includes the following topics:

- [Overview, 17](#)
- [Human Tasks, 17](#)
- [Probabilistic Models, 18](#)
- [Classifier Models, 19](#)
- [Hadoop, 20](#)
- [Web Services, 20](#)
- [Multithreading, 21](#)

Overview

You can configure several data quality components in addition to the transformations that define data analysis and enhancement operations.

Human Tasks

Informatica implements the ActiveVos engine to allow you to run workflows and resolve issues and conflicts manually with Human tasks. An administrator provides a database connection for the Human task metadata. If you encounter bottlenecks, you may need to increase the number of connections.

Workflow Orchestration Service Connection Pooling

The default value for the maximum number of Human task database connections is 15. This value may be sufficient for basic Human task processing. Increase the value to 64 for medium load and 128 for a very heavily used Human task interface. You can configure the value in the Workflow Orchestration Connection property on the Data Integration Service in the Administrator tool.

Data Integration Service Java Heap Size

The memory heap size allocated to the Data Integration Service is set as an advanced process property on the Data Integration Service. Set this value to a minimum of 1024 MB for Human tasks and to 3072 MB for a heavily-used Workflow Orchestration Service.

The following image shows the Maximum Heap Size property on the Data Integration Service in the Administrator tool:



For information on additional fine tuning on this value, contact Informatica Global Customer Support.

Human Task Load Balancing

When you configure load balancing for Human tasks, consider the number of tasks that will be created and the number of rows per tasks. Configure your load balance criteria appropriately for a useful distribution of tasks and data. For example, you are unlikely to create 5,000 tasks with one row of data each or to create one task with 5,000 rows.

You can choose from the following options when configuring load balancing:

- By number of items.
 - Number of items per task. The number of tasks created will be determined by the amount of work to be reviewed.
 - Number of tasks. The specified number of tasks will be created, and the workload will be split across the tasks.
- By data value. The number of tasks created is unknown prior to run time.

Probabilistic Models

Probabilistic models are reference data files that identify the type of information in each value in a text string. Consider the information below when you add probabilistic models to a Data Quality installation.

Design-Time Considerations

When you add training data to compile the model, all rows are loaded in memory. The default heap size of 768 MB available to the Developer tool means you might have access to approximately 500,000 rows. To edit larger training data sets, increase the heap size.

To increase the heap memory available, update the `-Xmx768M` value in the `developer.ini` file. Tests indicate that 100,000 rows of data require 100 MB of memory.

Model Creation

The Content Management Service performs compilation operations for probabilistic models. As probabilistic model compilation is memory-intensive, you may need to increase the heap size available to the Content Management Service.

When you compile a model, check the Content Management Service logs at the following location for errors:

\$INFA_HOME/logs/[nodename]/services/ContentManagementService

If there is insufficient memory to compile the model, you will see an error like this:

```
com.informatica.cms.service.webapp.ContentManagementServiceServlet
$ContentManagementServiceDefaultUncaughtExceptionHandler uncaughtException
WARNING: uncaughtException in CMS - Java heap space
```

When you see such an error, the Content Management Service has insufficient memory to compile the model, and you must increase the java heap size for the process. You can increase the java heap size in the Administrator tool. Navigate to the Processes tab on the Content Management Service, and update the Maximum Heap Size advanced property.

You specify the heap size in megabytes, for example 2048M, or gigabytes, for example 2G. Note the syntax in this case, as mistakes are common. After you update the heap size, restart the Content Management Service.

The average model consumes approximately 50 MB in physical disk space.

Note: If you are not concerned about the probabilistic score output from a Labeler or Parser transformation, leaving the score port unconnected on the transformation will improve performance.

Classifier Models

Classifier models are reference data files that identify the most common type of information in long text strings.

Consider the information below when you add classifier models to a Data Quality installation.

Design-Time Considerations

When designing a Classifier model, the resources required on the Developer tool are similar to the resources required for probabilistic models. You should not need to increase the Java heap size attributed to the Java process unless you are editing hundreds of thousands of rows of data. If cases exist where you are editing such volumes, you may need to increase the heap size in the developer.ini file.

Model Creation

Classifier models are compiled under the Content Management Service in the same way as probabilistic models. The classifier compilation process does not require the amount of memory that the probabilistic models require. As a result, you do not need to increase the resources available to the Content Management Service in order to successfully compile classifier models.

Hadoop

In some cases, the default Java heap size is insufficient for execution of mappings on a Hadoop cluster. For example, if you push down a mapping that reads a probabilistic model, you may need to increase the heap size. If such an issue occurs, increase the default Java heap size to eliminate errors.

The following code fragment increases the Java heap size to 1024 MB:

```
$INFA_HOME/services/shared/Hadoop/conf/hadoopEnv.properties

# Extra Java Options.

infapdo.java.opts=-Djava.library.path=$HADOOP_NODE_INFA_HOME/services/shared/bin -Xmx1024M -
Djava.security.egd=file:/dev/./urandom
```

You may see a Java heap size error in the Jobtracker logs, which are typically available at <http://<NameNode>:50030/jobtracker.jsp>.

Web Services

A web service client can connect to an Informatica web service to access, transform, or deliver data. An external application or a Web Service Consumer transformation can connect to a web service as a web service client. You can create an Informatica web service in the Developer tool.

Consider the information below when you configure Data Quality for web services.

DTM Keep Alive Time

DTM is an abbreviation for Data Transformation Model. It identifies the area of memory where a mapping and its metadata are loaded. Normally, once the mapping has finished, this memory is discarded. In the case of web services, you can instruct your Informatica installation not to destroy this memory and make it available if another request is received.

The DTM Keep Alive Time value defines how long to keep the DTM running after it has dealt with a request. Configure the keep alive time as high as the available resources permit.

When a web service request is received and no DTM is available, Data Quality initializes a new DTM to deal with the request. The initialization process may mean loading reference tables, address reference data, or probabilistic models into memory. The process may take seconds to complete and considerably increase the response time. If all the above are preloaded to the DTM, the response time will be milliseconds as opposed to seconds.

You can configure the DTM Keep Alive Time value as a Data Integration Service property in the Administrator tool.

The following image shows the DTM Keep Alive Time property on the Data Integration Service:



You can also set a web service-specific keep alive time. This value takes priority over the value set on the Data Integration Service. To set the value, browse to the web service under the Applications tab of the Data Integration Service in the Administrator tool.

The following image shows the DTM Keep Alive Time property on the web service:

▼ Web Service Properties	
Startup Type	Enabled
Trace Level	INFO
Request Timeout	3600000
Maximum Concurrent Requests	10
Sort Order	Binary
Enable Transport Layer Security(TLS)	false
Enable WS-Security	false
Optimization Level	2
DTM Keep Alive Time	-1
SOAP Output Precision	200000
SOAP Input Precision	200000

Note: The DTM Keep Alive Time value is specified in milliseconds unless you set a negative integer value. Set a negative integer in the DTM Keep Alive Time property for the web to specify that the web service will read the property value from the Data Integration Service.

Maximum Execution Pool Size

The Maximum Execution Pool Size property defines the maximum number of mappings that can run concurrently in the Data Integration Service. The larger the number, the more resources the Data Integration Service will require.

Maximum Concurrent Requests

The Max Concurrent Requests property defines the maximum number of concurrent HTTP or HTTPS requests that the Data Integration Service process can accept. The larger the number, the more resources the Data Integration Service will require.

Maximum Backlog Requests

The Max Backlog Requests property defines the maximum number of HTTP or HTTPS requests that can reside on the queue for the Data Integration Service process before the requests are rejected.

Note: Once you have developed your web service, you can turn off logging on the web service. This will increase performance as no logs will be written for each request.

Multithreading

You can run mappings in a multi-threaded or parallel manner. The Execution Options on the Data Integration Service define the maximum number of parallel mappings that the service can run.

The following image shows the maximum parallelism option on the Data Integration Service:

Execution Options	
Launch Job Options	In separate local processes
Maximum Execution Pool Size	10
Maximum Memory Size	0
Maximum Parallelism	15
Hadoop Kerberos Service Principal Name	
Hadoop Kerberos Keytab	
Home Directory	/
Temporary Directories	/tmp
Cache Directory	/cache
Source Directory	/source
Target Directory	/target
Rejected Files Directory	/reject
Informatica Home Directory on Hadoop	/opt/informatica
Hadoop Distribution Directory	/opt/informatica/services/shared/hadoop/ Hortonworks_2.2
Data Integration Service Hadoop Distribution Directory	/../services/shared/hadoop/ Hortonworks_2.2

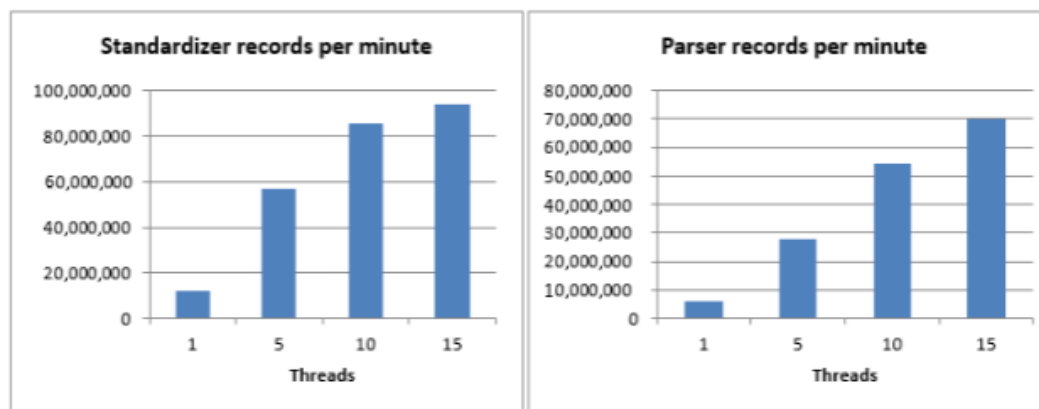
Parallelism also applies to the transformations within a mapping. Set the maximum permitted parallelism within a given mapping as a run-time property on the mapping.

The following image shows the maximum parallelism option on the mapping:

The screenshot shows the 'Properties' window for a mapping, with the 'Run-time' tab selected. Under the 'Execution Environment' section, the 'Maximum Parallelism' property is highlighted, showing a value of 15. Below this, a description states: 'Maximum number of parallel threads that process a single pipeline stage.'

All data quality transformations can be multithreaded except for the Exception, Association, Classifier, and Consolidation transformations. You can configure a Decision transformation to be partitionable.

The following graphs show the increase in throughput that you can achieve by enabling partitioning on a Standardizer and Parser transformation:



Similar performance increases are observed for other data quality transformations.

Consider the following rules and guidelines for parallelism:

- The number of execution instances that you set on a Match transformation or Address Validator transformation must not exceed the maximum parallelism values that you set on the Data Integration Service or on the mapping that contains the transformation.
- If you set the maximum parallelism value on a mapping to Auto, the mapping uses the maximum parallelism value on the Data Integration Service. This may result in diminished performance, depending on the number of mappings that run concurrently.