SAP Data Services

Document Version: 4.2 Support Package 7 Patch 1 (14.2.7.1) – 2017-01-24

Sizing Guide



Content

1	About this guide	3
2	About SAP Data Services	4
2.1	SAP Data Integrator	4
2.2	SAP Text Data Procesing	4
2.3	SAP Data Quality	5
3	SAP Data Services architecture	6
4	Factors that influence performance of SAP Data Services	7
4.1	Performance of data integration	7
4.2	Performance of data quality	8
4.3	Performance of text data processing	8
5	Server hardware and system landscape	9
6	Batch sizing guidelines	10
6.1	Scenario 1: Loading customer data from Oracle to SAP HANA	11
6.2	Scenario 2: Loading customer data from a flat file to SAP HANA	12
6.3	Scenario 3: loading customer dimension with validation operations	13
6.4	Scenario 4: loading customer dimension with table comparison operations	14
6.5	Scenario 5: Joining tables to load a fact table	14
7	Real-time sizing guidelines	16
7.1	Simple scenario: cleansing US addresses	18
7.2	Moderate scenario: cleansing world-wide addresses	18
7.3	Complex scenario: identifying duplicates	19

1 About this guide

This guide explains how to perform sizing for SAP Data Services.

This document demonstrates how someone might perform sizing of an SAP Data Services system. The methodology and sample walkthroughs offered here are examples of the tasks and thinking involved. The performance and functioning of an actual system may vary for many reasons. The examples offered here should not be considered a guarantee of success of a particular deployment.

You should work with your SAP account representative to ensure you get the advice of a sizing expert when making final decisions about your Data Services deployment.

2 About SAP Data Services

SAP Data Services delivers a single enterprise-class solution for data integration, data quality, data profiling and text data processing.

SAP Data Services delivers a single enterprise-class solution for data integration, data quality, data profiling and text data processing that allows you to integrate, transform, improve and deliver trusted data to critical business processes. It provides one development UI, metadata repository, data connectivity layer, run-time environment and management console – enabling IT organizations to lower total cost of ownership and accelerate time to value. With SAP Data Services, IT organizations can maximize operational efficiency with a single solution to improve data quality and gain access to heterogeneous sources and applications.

2.1 SAP Data Integrator

SAP Data Integrator allows you to integrate data form multiple differing sources.

- Easy-to-configure transforms for typically complex tasks like slow changing dimensions, hierarchy flattening, etc.
- Everything you need to build large jobs including error handling, dependency handling, and restart-ability
- Extensive operational statistics
- Rich connectivity to many sources and targets most using the vendor's native format for maximum performance
- Easy-to-use parallelization and performance optimization options
- Functionalities to simplify daily operations and project hand-over like web based management console, autodocumentation features and impact lineage information

2.2 SAP Text Data Procesing

SAP Text Data Procesing enables you to analyze text in detail.

- Analyzes text and automatically identifies and extracts entities including people, dates, places, organizations and so on, in multiple languages
- Looks for patterns, activities, events, and relationships among entities and enables their extraction
- Goes beyond conventional character matching tools for information retrieval, which can only seek exact matches for specific strings; understands the semantics of words
- Supports extraction in 31 different languages
- Supports not only text, HTML, and XML but binary document formats such as PDF and Microsoft Word
- Allows you to specify your own list of entities in a custom dictionary, which enables you to store entities and manage name variations; known entity names can be standardized using a dictionary
- Write rules to customize extraction output, although pre-defined rules are provided to support sentiment analysis, enterprises, and the public sector

- Broad, heterogeneous application and system support for both SAP and non-SAP sources and targets
- High performance and scalability with software that can meet high volume needs through parallel processing, grid computing, and bulk data loading support

2.3 SAP Data Quality

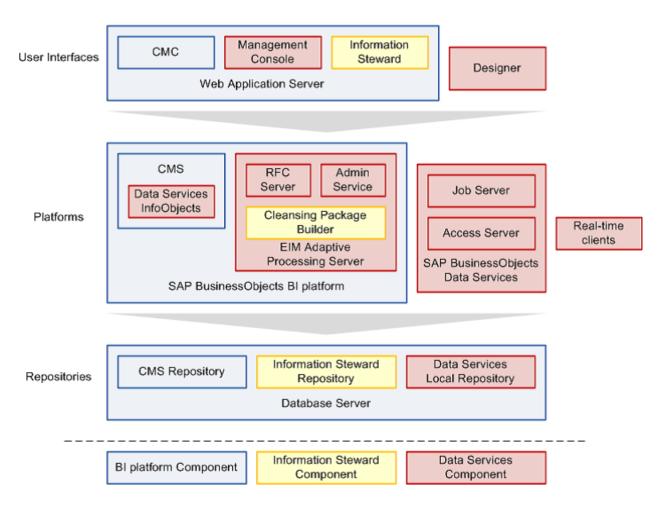
SAP Data Quality enables you to correct, standardize, and eliminate duplicates from your data.

- Data quality dashboards that show the impact of data quality problems on all downstream systems or applications
- Ability to apply data quality transformations to all types of data, regardless of industry or data domain such as structured to unstructured data as well as customer, product, supplier, and material information
- Intuitive business user interfaces and data quality blueprints to guide you through the process of standardizing, correcting, and matching data to reduce duplicates and identify relationships
- Comprehensive global data quality coverage with support for over 230 countries
- Comprehensive reference data
- Broad, heterogeneous application and system support for both SAP and non-SAP sources and targets
- Prepackaged native integration of data quality best practices for SAP environments
- Optimized developer productivity and application maintenance through intuitive transformations, a centralized business rule repository, and object reuse
- High performance and scalability with software that can meet high volume needs through parallel processing, grid computing, and bulk data loading support
- Flexible technology deployment options, from an enterprise platform to intuitive APIs that allow developers quick data quality deployment and functionality

3 SAP Data Services architecture

This topic shows the SAP Data Services architecture.

This diagram illustrates how SAP Data Services components fit in with other software in the SAP portfolio.



More details about the architecture of SAP Data Quality Management and SAP Data Services can be found in the SAP Data Services Administrator's Guide.

4 Factors that influence performance of SAP Data Services

Many common factors can influence the performance of SAP Data Services.

- Access to source and targets The bandwidth to the source and target can affect how fast data can be passed through the dataflow
- Availability of additional RAM If caching is needed, allocating enough free RAM within the system will speed up the dataflow, not only to cache lookup data but also reference data for Data Quality transforms
- Configuration and system landscape This sizing guide was created with SAP Data Services installed on the target database system. Source RDBMS was located on a separate machine. Competing applications running multiple resource intensive applications may cause competition for the resources and reduce the throughput for an individual job
- **Degree of Parallelization (DOP)** The DOP setting can greatly influence performance when the appropriate hardware is utilized; increasing this setting will generally increase throughput
- **Grid computing** Using grid computing on multiple servers can process different elements of the same job in parallel; a server group can be created to group several job servers, and Data Service's load balancing algorithm will make sure the server with the lowest workload will execute the processes that are spawned by the job
- Parallel file reader The "Parallel Process Threads" option allows you to specify the number of parallel threads that will be used to convert the text string read from the flat file to data organized in columns based on the schema defined in the file format

4.1 Performance of data integration

These factors can incluence the performance of data integfration.

- Loader method Each loading method has pros and cons, depending on how the database vendor implemented it. Depending on the databases (and versions), different loader options can have dramatic differences in performance, regular load, Bulkloading, AutocorrectLoad.
- **Transactional loaders** Loading data in one transaction means the dataflow cannot use parallel sessions to speed up the loading
- Lookup and join settings The amount of data to be processed versus the size of the lookup table. SAP Data Services lets you choose the best lookup strategy; choose carefully.
- **Heterogeneous sources or all in one database** If all data is in one database or a database link exists between the databases, the SAP Data Services optimizer has more options so it can decide to delegate parts or all processing to the database.

4.2 Performance of data quality

These factors can incluence the performance of data quality.

- Complexity of processing A data quality transform can do varying degrees of simple or complex processing based on the options set for the transform; generally, the more complex the processing, the more hardware resources needed
- Location of reference data Several of the cleansing transforms use reference data located on the file system and can be I/O or network dependent; the speed of this access will affect the performance of these transforms

4.3 Performance of text data processing

Your document's characteristics--including the format, length, and density--can impact the performance of text data processing.

- Format XML and HTML require de-tagging before processing the text which has more overhead than processing text directly. Additionally, converting a binary document into a textual representation during processing has overhead
- Length Longer input documents require more processing time
- Density More density and fact-rich input documents require more processing time
 In addition to the above factors, using one or more rules to customize extraction may require more processing time.

5 Server hardware and system landscape

Your server hardware and system landscape have a direct effect on system performance.

Your server hardware and system landscape have a direct effect on system performance. We used the system landscape outlined in the tables below to generate the results shown in this guide. Note also that SAP Data Services, source, and target RDBMS were installed on the same server to simulate a 10G network environment.

Hardware

Table 1:

Mode:	Z820 CPU: Intel(R) Xeon(R) CPU E5-2670 v2 @ 2.50GHz
OS:	SUSE Linux Enterprise Server 11 (x86_64)
DISK:	SSD 1.3 TB
MEMORY:	RAM 512GB
CPU CORE:	40

Software

Table 2:

Data Services	4.2 SP4
BOE	4.1 SP4

Database

Table 3:

SAP HANA	1.0 Rev 90
Oracle	11G
SAPERP	7.0

Configuration

SAP Data Services, source, and target RDBMS are installed on the same server to simulate a 10G network environment.

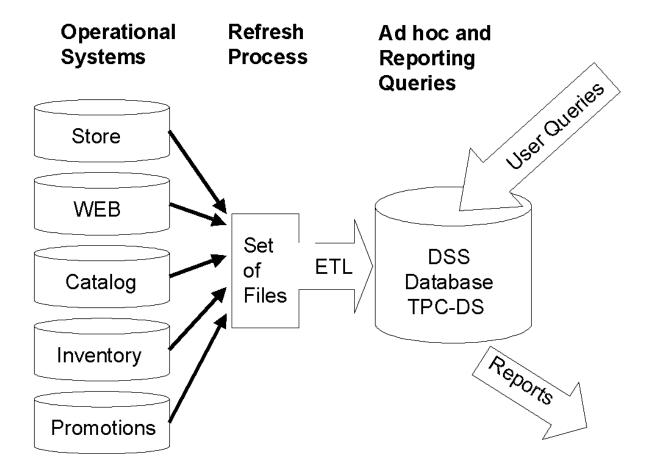
6 Batch sizing guidelines

This topic provides an overview of batch sizing guidelines.

Our goal is to provide data integration performance numbers for real-life scenarios, rather than focusing on record-breaking performance numbers in an isolated and unrealistic test environment. This approach will enable you to compare different approaches and results with your environment, and use the numbers in this document to estimate the time needed to meet your specific data loading requirements, which can be the foundation for a sizing study.

Data sets

For the source data we used the TPC-DS data model (www.tpc.org) which represents a real life enterprise resource planning-like (ERP) source system; representative data is generated automatically for all tables.



	JSTOMER(hana_ds.PERF_USER)	
[C_CUSTOMER_SK	int
[C_CUSTOMER_ID	varchar(16)
[C_CURRENT_CDEMO_SK	int
	C_CURRENT_HDEMO_SK	int
[C_CURRENT_ADDR_SK	int
[C_FIRST_SHIPTO_DATE_SK	int
[C_FIRST_SALES_DATE_SK	int
[C_SALUTATION	varchar(10)
[C_FIRST_NAME	varchar(20)
[C_LAST_NAME	varchar(30)
[C_PREFERRED_CUST_FLAG	varchar(1)
[C_BIRTH_DAY	int
[C_BIRTH_MONTH	int
[C_BIRTH_YEAR	int
[C_BIRTH_COUNTRY	varchar(20)
[C_LOGIN	varchar(13)
[C_EMAIL_ADDRESS	varchar(50)
	C_LAST_REVIEW_DATE	varchar(10)

Test concept overview

For the test scenarios we've chosen the example of loading a data warehouse from a transactional system. However, most examples can be applied to other data migration scenarios as well.

Data migration use cases

One of the primary use cases for Data Services is to move data from one application to another. Migration projects are relevant for all types of companies, including those with SAP applications, non-SAP applications, and upgrading: any company who needs to migrate data from one application to another.

6.1 Scenario 1: Loading customer data from Oracle to SAP HANA

Loading customer data from Oracle to SAP HANA.

The table below shows the results of loading a customer table from an Oracle database to an SAP HANA database using 8 loaders and 8 partitions.

Table 4:

# of rows	Size of data	Throughput (GB/ Hour)	CPUs	Memory
2 Million	0.25 GB	149.49 GB	8 cores	4 GB

# of rows	Size of data	Throughput (GB/ Hour)	CPUs	Memory
10 Million	~1.38GB	187.67 GB	8 cores	4 GB
50 Million	~6.92GB	188.45 GB	8 cores	4 GB
100 Million	~ 13.68GB	160.19 GB	8 cores	4 GB
500 Million	~ 68.3 GB	148.91 GB	8 cores	4 GB

Because the database is the bottleneck, the CPU is not used 100% on the Data Services server. Tuning Data Services will not bring a lot of performance improvement; throughput from the database side needs to be enhanced first (for example, by adding more disks for faster I/O). In general, increasing the number of parallel file threads and loaders will improve the overall performance throughputs. In general, using parallel execution by increasing the number of parallel file threads and loaders will improve the overall performance throughputs.

6.2 Scenario 2: Loading customer data from a flat file to SAP HANA

Loading customer data from a flat file to SAP HANA.

The table shows the results of loading data from a flat file to an SAP HANA database using 8 loaders.

Table 5:

# of rows	Size of data	Throughput (GB/ Hour)	CPUs	Memory
2 Million	0.25 GB	154.64 GB	8 cores	4 GB
10 Million	~1.38GB	276.86 GB	16 cores	4 GB
50 Million	~6.92GB	296.25 GB	16 cores	4 GB
100 Million	~ 13.68GB	278.58 GB	16 cores	4 GB
500 Million	~ 68.3 GB	282.09 GB	16 cores	4 GB

In general, increasing the number of parallel file threads and loaders will improve the overall performance throughputs. The table below illustrates the improvements for SAP Data Services engine performance using different configurations. For example, if we increase the number of parallel threads from 4 to 16, you can see SAP Data Services improves performance to 315.86 GB per hour in comparison to 171.19 GB per hour. Keep in mind that the more loaders that are used, the more memory SAP Data Services will require.

Table 6:

# of File threads	# of Loaders	# of rows	Size of data	Throughput (GB/Hour)	CPUs	Memory
4	4	100 Million	12.4GB	171.19 GB	8 cores	4 GB
8	8	100 Million	12.4GB	278.58 GB	16 cores	8 GB
16	16	100 Million	12.4GB	315.86 GB	20 cores	12 GB
4	4	500 Million	62.3 GB	162.39 GB	8 cores	4 GB
8	8	500 Million	62.3 GB	282.09 GB	16 cores	8 GB
16	16	500 Million	62.3 GB	355.97 GB	28 cores	12 GB

6.3 Scenario 3: loading customer dimension with validation operations

In this scenario, we load the customer table from Oracle to SAP HANA, where the validation transforms need to be executed.

In our example, we defined three business rules stating customer first and last name cannot be NULL and the email address that is made up of a local part, an @ symbol, then a domain such as john.smith@example.com.

Table 7:

Operations	Execution time (sec)	Throughput (GB/ Hour)	CPUs	Memory
No validation during loading	22	156 GB	2 cores	4GB
Validation during load- ing	32	108 GB	2 cores	4GB

To get an idea about the validation transform's performance, we executed two tests. For the first test, we loaded the customer table directly from the source database (Oracle) to the target database (SAP HANA). In the second test, we used the same customer table, but added a validation transform in the dataflow to validate the business rules before loading the data into the target. As the result shows, SAP Data Services can still perform 108GB per hour and did not introduce significant bottleneck for a more complex data load.

6.4 Scenario 4: loading customer dimension with table comparison operations

This scenario loads a customer dimension with table comparison operations.

For delta loads, SAP Data Services can tightly integrate with CDC techniques provided by your database (such as Oracle CDC) or SAP Replication Server for optimal performance. In case this is not available on your source system, SAP Data Services' built-in Table Comparison transform can be used to detect changes by comparing the source with the target table. In our 2,000,000 record customer table, we achieve a throughput of 14 GB/hour for the delta load while comparing each source record with the target and keeping history for changed records.

Table 8:

	Tuning parame- ters	Execution time (sec)	Throughput (GB/ Hour)	CPUs	Memory
Initial load		22	156 GB	2 cores	4GB
Delta load	row-by-row, DOP = 2	248	14 GB	2 cores	4GB
Delta load	row-by-row, DOP =	118	30GB	2 cores	4GB
Delta load	Cache, DOP = 4	40	86 GB	2 cores	4GB
Delta load	Sorted input, DOP = 2 or 4	25	137GB	2 cores	4GB

The above table compares the results used row-by-row select and cached comparison table. As the result shows, the row-by-row selects each row from the source, and calls a query on the target to verify if it's a new, an existing or a changed row. This is usually the slowest method. In contrast, the cached comparison table technique compares columns from the target table that are read into memory first and the comparison is done in memory time. This technique is usually performed very fast for smaller tables, but might be slower for larger tables (> 2GB) as paging may occur. If the memory-consuming operations in your data flow exceed the available memory, use pageable cache to complete the operation. To tune performance for the Table Comparison transform, you may consider sorting the input data in both source and target. This will be the fastest approach as you can see in above table.

6.5 Scenario 5: Joining tables to load a fact table

This scenario shows joining tables to load a fact table.

In a typical data warehouse project, a retailer may wish to pull a sales report for a particular store during a specific period of time. The retailer can do this by retrieving information from the fact table that relates to those transactions meeting the specific criteria. For example, you may join the store sales, date and customer

dimension tables, and generate a data warehouse key for each new record before loading the fact table into the target database.

Table 9:

Source Table	Store_Sales	Date	Customer
# of rows	287,997,204	73,049	2,000,000
Total Table Size (GB)	37.88	0.009	0.249

The join can be performed on the database or in the SAP Data Services engine. The common trend is to try to push down as much as possible to the database for performance optimization. In the first scenario, all tables are stored in the same database. Therefore, SAP Data Services can leverage the power of the database server and do a full push-down to the source database.

Table 10:

# of rows	Size of data	Job Completion	CPUs	Memory
5,244,914	313.65 GB	90 seconds	2 cores	4GB

However, we cannot assume all tables will always be stored in the same database. In case the tables reside on different databases, joining in the SAP Data Services engine will be the only option.

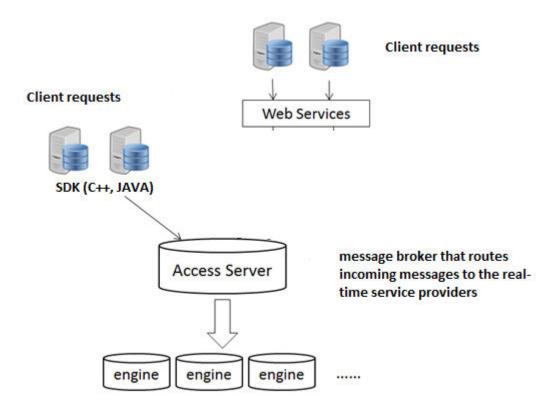
Table 11:

# of rows	Size of data	Job Completion	CPUs	Memory
5,244,914	313.65 GB	30 minutes	2 cores	4GB

7 Real-time sizing guidelines

This topic provides an overview of real-time sizing guidelines.

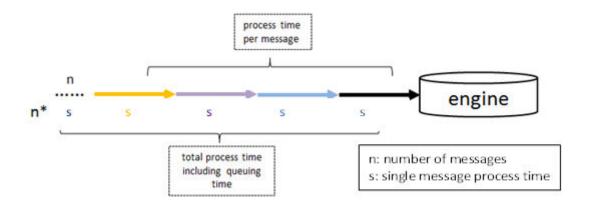
SAP Data Services provides Web Services, JMS and SDK library (C++ and Java API's) to programmatically pass data to SAP Data Services for real-time data movement, transformation, and retrieval. Real-time processing has slightly more overhead than batch processing because of the work required for processing client requests and distributing them to the job servers via the access servers.



Sizing calculation

When sizing for transactional processing, it's important to consider response time requirements, estimated peak transactional throughput needs, and the number of potential concurrent client requests that needs to be supported.

In addition, you may consider increasing the number of corresponding service providers as the number of concurrent messages increased for performance optimization. This is because SAP Data Services uses the round-robin scheduling algorithm to split the incoming messages to each process engine in equal portions and in circular order. The figure below illustrates message queuing time in SAP Data Services has increased when the incoming message continuous to grow.



As a general guideline, sizing for transactional processing can be calculated as follows:

- 1. Identify the engine processing time. Send a single message to SAP Data Services and check the maximum processing time. The queuing time should be zero.
- 2. Identify the number of required engines. As the concurrent messages continues to grow, it is best to configure SAP Data Services with additional engines to optimize the message processing time. Observe the average processing time to see if it meets the desired average performance time.
- 3. Identify the minimum system resource requirement. System resource requirement needs to consider other processes including job server, web server, engine processes recycling, and BOE/IPS CMS process. Therefore, you must allocate enough system resources for your SAP Data Services installation when performing your sizing exercise. For minimum hardware requirements to run SAP Data Services products in one or multiple landscapes, see the *Product Availablility Matrix* (PAM) available at: https://apps.support.sap.com/sap/support/pam. By default, the size of an empty IS/DS repository is insignificant. However, the size of the repository will expand based on the increase of the number of objects within the repository. Each job and datastore that are added, plus the runtime statistics captured within al_statistics, al_history, etc. all increase the size of the repository. The number of objects stored, and the longer that runtime statistics are retained, the bigger the size of the repositories. Here is a snapshot of the minimum disk space to install Data Services with miniBOE (IPS) with an empty repository.

Table 12: Disk space in empty DS/IS repository

Component	Disk space
SAP BusinessObjects folder: includes: IPS, Tomcat, DS, Sqlanywhere, etc	~16-18GB
Default installation with all languages	~23GB
Empty DS Repository	~25MB
Empty IS Repository	~65MB

4. Plan for more system resources when an additional engine is configured. To calculate the total memory and CPU consumption, multiply the memory and CPU per engine with the number of engines configured. For example, if m = number of engines, Total RAM consumption = m*memory consumption per engine, and Total CPU consumption = m*CPU utilization per engine.

Test concept overview

Typical real-time messages include information required to implement a particular business operation and to produce an appropriate response. For example, an address-entry system is integrated with SAP Data Services via Web Services or API's in a call center, where cleansing is necessary at the point of entry for customer data.

7.1 Simple scenario: cleansing US addresses

This scenario parses, standardizes, and corrects US address data.

Suppose a message contains information required to create a new customer account. The message content might be as simple as the customer name along with the contact information. The first step is to send the message to SAP Data Services to cleanse customer data by performing data parsing, standardization and correction, and receive the corrected results back from SAP Data Services to the client application.

Table 13:

Number of con- current client re- quests	Number of engines	Average engine processing time	# of messages processed (in sec- ond)	Memory con- sumption for DS engine(s)	Core
1	1	< 10 ms	103.55	1 GB	2
5	1	< 20 ms	383.05	1 GB	4
5	5	< 10 ms	607.90	5 GB	6
10	1	< 20 ms	741.93	1 GB	6
10	5	< 10 ms	1856.98	5 GB	10
50	5	< 30 ms	2176.17	5 GB	14
50	10	<20 ms	3188.93	10 GB	16

As a general rule of thumb, it is necessary to add more engines as the concurrent clients increased to minimize the queuing time for better performance.

7.2 Moderate scenario: cleansing world-wide addresses

This scenario parses, standardizes, and corrects world-wide address data.

The table below shows the results of parsing, standardizing, and correcting world-wide address data.

Table 14:

Number of con- current client re- quests	Number of engines	Avg processing time	# of messages processed (in sec- ond)	Memory con- sumption for DS engines	Core
1	1	< 10 ms	88.53	1 GB	2
5	1	< 20 ms	191.73	1 GB	4
5	5	< 10 ms	572.40	5 GB	6
10	1	< 40 ms	201.20	1 GB	6
10	5	< 10 ms	936.27	5 GB	10
50	5	< 30 ms	1148.93	5 GB	14
50	10	< 20 ms	1521.73	10 GB	16

7.3 Complex scenario: identifying duplicates

This scenario identifies duplicates based on customer data.

The table below shows the results of identifying duplicates based on customer data.

Table 15:

Number of con- current client re- quests	Number of engines	Avg processing time	# of messages processed (in sec- ond)	Memory con- sumption for DS engines	Core
1	1	<10 ms	75.60	1 GB	2
5	1	< 20 ms	201.45	1 GB	4
5	5	< 10 ms	544.05	5 GB	6
10	1	< 40 ms	194.55	1 GB	6
10	5	< 10 ms	1074.45	5 GB	10
50	5	< 30 ms	970.35	10 GB	16
50	10	< 20 ms	1340.25	10 GB	16

Based on the numbers above, we can conclude that the processing overhead for handling requests in SAP Data Services outweighs the variances in performance of our SAP Data Quality transforms, so whether you're doing cleansing, geocoding, matching, etc. the overall sizing requirements are similar.

Important Disclaimers and Legal Information

Coding Samples

Any software coding and/or code lines / strings ("Code") included in this documentation are only examples and are not intended to be used in a productive system environment. The Code is only intended to better explain and visualize the syntax and phrasing rules of certain coding. SAP does not warrant the correctness and completeness of the Code given herein, and SAP shall not be liable for errors or damages caused by the usage of the Code, unless damages were caused by SAP intentionally or by SAP's gross negligence.

Accessibility

The information contained in the SAP documentation represents SAP's current view of accessibility criteria as of the date of publication; it is in no way intended to be a binding guideline on how to ensure accessibility of software products. SAP in particular disclaims any liability in relation to this document. This disclaimer, however, does not apply in cases of willful misconduct or gross negligence of SAP. Furthermore, this document does not result in any direct or indirect contractual obligations of SAP.

Gender-Neutral Language

As far as possible, SAP documentation is gender neutral. Depending on the context, the reader is addressed directly with "you", or a gender-neutral noun (such as "sales person" or "working days") is used. If when referring to members of both sexes, however, the third-person singular cannot be avoided or a gender-neutral noun does not exist, SAP reserves the right to use the masculine form of the noun and pronoun. This is to ensure that the documentation remains comprehensible.

Internet Hyperlinks

The SAP documentation may contain hyperlinks to the Internet. These hyperlinks are intended to serve as a hint about where to find related information. SAP does not warrant the availability and correctness of this related information or the ability of this information to serve a particular purpose. SAP shall not be liable for any damages caused by the use of related information unless damages have been caused by SAP's gross negligence or willful misconduct. All links are categorized for transparency (see: http://help.sap.com/disclaimer).

