Draft Document for Review November 7, 2018 6:19 am REDP-5520-00

Redbooks
ibm.com/redbooks

Best Practices Guide for Databases on IBM FlashSystem

Jagadeesh Papaiah

Storage

IBM

Redpaper

Best Practices Guide for databases on IBM FlashSystem

The purpose of this document is to provide best practice guidelines to design and implement IBM® FlashSystem storage for database workloads. The recommended settings and values are based on lab testing, Proof of Concept (PoC) and experience drawn from Customer implementations. Recommendations presented in this document are applicable to most production database environments to increase performance of IO and availability, however more considerations might be required while designing, configuring and implementing storage for extreme transactional, analytical and Database cluster environments.

Customers have been migrating database storage to FlashSystem largely due to low latency performance of IBM FlashSystem® family of Storage. Using FlashSystem our Customers have been able to achieve, low latency for queries and transactions from milliseconds to microseconds, realize a multifold increase in application level transactions per second, increase CPU efficiency and reduce database licensing costs. Recent additions of data reduction technologies to FlashSystem further adds to overall TCO benefits. All FlashSystem models now offer compression which can reduce database storage by 40 - 80% depending on database software.

In addition to Best Practices described in this document FlashSystem Worldwide Solutions Engineering Team can further assist Customers with performing analysis of current database workloads for FlashSystem benefits, perform PoCs at our Labs and help with implementation.

Oracle

Oracle remains our most popular database on FlashSystem in terms of size of the databases and the resulting low latency requirements. For Oracle databases FlashSystem average latency ranges from 300 microseconds to 1 millisecond depending on FlashSystem model and database workload type variations.

Table 1 shows the recommended init.ora values, ASM and logfile considerations to achieve a balanced performance, also refer to OS consideration sections and Storage layout for settings corresponding to Operating system and FlashSystem model.

Table 1 Shows the database for Oracle systems settings

Parameter	Default setting	Recommendations	Description
FILESYSTEMIO_OPTIONS	Varies by database and OS	SETALL	Set all enables both direct and asynch IO
Block size	8KB, range 2K-32K	Not modifiable after DB creation, 8K optimal for most DBs	Large block size for LOBs and set at tablespace level
DB_FILE_MULTIBLOCK_READ_COUNT	Default value corresponds to the maximum I/O size and is platform dependent	32 optimal for FlashSystems, At 32 MBR and Blksize 8KB, avg read scan size issued to 256KB sequential for table scans, see table below for testing results	Specifies the maximum number of blocks read in one I/O operation during a sequential scan
Redo log size	File size is determined during creation of redo log files	4GB per logfile and minimum of 4 groups and for HA 2 members placed on separate storage array	Larger logfile reduces the number log switches
Redo logfile block size	512 Bytes	4K blksize, set "_disk_sector_size_override"=TRUE, to add logfile with 4K blksize	On FlashSystem 4k block size is optimal and reduces 'log file synch' waits
FAST_START_MTTR_TARGET	3600 seconds	Parameter should be adjusted and tested to meet RTO requirements	Specifies number of seconds the database takes to perform crash recovery
Log buffer size	5 MB to 32 MB,	32 - 64MB,	depending on the size of the SGA, CPU count, and whether the operating system is 32-bit or 64-bit
ASM Disk Redundancy	No default, options External = 1x, Normal = 2x, High = 3x copies	External	For High availability create disk group as 'Normal' and mirror disks across two arrays

Parameter	Default setting	Recommendations	Description
ASM Allocation Unit - AU size	1 MB	4 or 8MB for A9000, 4 MB for V9000 and FS900	4MB is optimal for most databases

On databases with more DSS/analytic type workloads, significant amount of table scans are is sued by Oracle database which results in large block sequential reads. Sequential reads to FlashSystem might be further tuned for optimization, following table illustrates query response time differences with varying multi block read count. Based on our lab testing using HammerDB TPCH schema tables, a combination of 8K blksize and 32 multiblock read count achieved the lowest response time.

Table 2 shows the query response time differences based on SQL queries against HammerDB TPCH schema tables. Negative query response time % differences are shown for varying block size MBR combination.

Table 2 Query response time:

ORA Block size KB	MBR Count	Query Response Time Differences	Read Scan Block Size KB
8	128	-25%	1024
8	64	_2%	512
8	32		256
8	16	-4%	128
8	8	-8%	64
8	4	-23%	32

Oracle I/O Calibration

Consider using Oracle provided stored procedure DBMS_RESOURCE_MANAGER.CALIBRATE_IO for IO calibration. Note this is optional and it is not required for FlashSystem implementation.

The I/O calibration feature of Oracle Database enables you to assess the performance of the storage subsystem, and determine whether I/O performance problems are caused by the database or the storage subsystem. Unlike other external I/O calibration tools that issue I/Os sequentially, the I/O calibration feature of Oracle Database issues I/Os randomly using Oracle data files to access the storage media, producing results that more closely match the actual performance of the database. This procedure issues an I/O intensive read-only workload, made up of one megabyte of random of I/Os, to the database files to determine the maximum IOPS (I/O requests per second) and MBPS (megabytes of I/O per second) that can be sustained by the storage subsystem.

The I/O calibration occurs in two steps:

- In the first step of I/O calibration with the DBMS_RESOUCE_MANGER.CALIBRATE_IO procedure, the procedure issues random database-block-sized reads, by default, 8 KB, to all data files from all database instances. This step provides the maximum IOPS, in the output parameter MAX_IOPS, that the database can sustain. The value MAX_IOPS is an important metric for OLTP databases. The output parameter MAX_LATENCY provides the

average latency for this workload. When you need a specific target latency, you can specify the target latency with the input parameter MAX_LATERNCY specifies the maximum tolerable latency in milliseconds for database-block-sized IO requests).

- The second step of calibration using the DBMS_RESOUCE_MANGER.CALIBRATE_IO procedure issues random, 1 MB reads to all data files from all database instances. The second step yields the output parameter MAX_MBPS, which specifies the maximum MBPS of I/O that the database can sustain. This step provides an important metric for data warehouses.

The calibration runs more efficiently if the user provides the number of physical disks input parameter, which specifies the approximate number of physical disks in the database storage system.

Executing procedure DBMS_RESOURCE_MANAGER.CALIBRATE_IO (<DISKS>, <MAX_LATENCY>, iops, mbps, lat);

Input values for FlashSystem DISKS = 12 or 24 and MAX_LATENCY = 10 or 20

Outputs maxiops, latency and maxmbps.

Caution: Due to the overhead from running the I/O workload, I/O calibration should only be performed when the database is idle, or during off-peak hours, to minimize the impact of the I/O workload on the normal database workload.

Microsoft SQL Server

SQL Server implementations on FlashSystem are largely using VMWare virtual machines, recommendations apply both bare metal and virtual machines. Additional considerations are listed under VMware considerations. Table 3 shows the parameters for SQL server.

Table 3 Parameters for Microsoft SQL Server

Parameter	Default setting	Recommendations	Description
Page size	8KB	Not modifiable	Disk I/O operations are performed at the page level
Extent Size	64KB	Not modifiable	Extent is eight physically contiguous pages, databases have 16 extents per megabyte.
MAXDOP	0	8 or specify the maximum number of processor cores that can be used by a single query execution.	Setting MAXDOP to 0 allows SQL Server to use all the available processors up to 64 processors.
Log files	1 logfile	Use separate drive for logs and use dedicated volumes for log files	IOs to logfile are primarily writes

Parameter	Default setting	Recommendations	Description
TempDB	1 datafile	Multiple, 1 datafile/cpu or core, and pre-size	On databases with significant sorts, multiple files and dedicated luns improves performance
Data files	1 datafile	1 data file (per filegroup) for each CPU on the host server, and pre-size	For large databases creating larger volumes on A9000 and multiple volumes 8-16 on V9000 improves performance
Backup - BUFFERCOUNT	Varies	32, commands in queue	Option can be set at SQL command level or at tools level
Backup - MAXTRANSFERSIZE	1 MB	2-4 MB	Option can be set at SQL command level or at backup tools level

Additional considerations for SQL Server

- UPDATE STATISTICS procedure WITH FULLSCAN, ALL or COLUMNS options for IO intensive databases after batch or load operations, by default query optimizer updates statistics as necessary to improve the query plan.
- Avoid mixing SQL and Exchange data

Db2

Table 4 shows the recommendations that apply to Db2® Linux, UNIX, Windows (LUW) and does not apply to Db2 on z/OS®.

Table 4 Db2 Linux, Unix and Windows recommendations

Parameter	Default setting	Recommendations	Description
Page Size	4K - 32K parameter contains the value that was used as the default page size when the database was created	4K for default	4K optimal for OLTP and 16 - 32K for analytics and LOB set at tablespace level
dft_extent_sz	32 pages	Use default size for all tablespaces	consider adjusting prefetch size for tablespaces with larger page size
dft_prefetch_sz	Automatic	Default good enough for most tablespaces	for tablespaces created with 32K page size consider lowering prefetch size to 16
Tablespace management	Automatic if managed by clause is not specified or specified as 'Automatic' during tablespace creation	Automatic	Use Automatic and avoid using SMS and DMS tablespace as they will be deprecated in future versions
OVERHEAD, DEVICE READ RATE	6.725ms, 100MB/s	Default	Defaults are ok for most databases and consider changing overhead to 1ms for high OLTP. If the Db2 database was upgraded from versions earlier to 10.1, then the existing tablespaces retain the overhead and device read rate attributes for that storage group which is set to undefined.

Operating System Considerations

This section shows the operating system settings optimal for FlashSystem based testing, Customer implementations and corresponding Vendor recommendations for Flash based Storage systems.

Linux

Table 5 shows the Linux parameters.

Table 5 Linux parameters

Parameter	Recommendations	Description
Block device scheduler	Noop or deadline, deadline for Oracle ASM disks and VMware	IO scheduler for Linux kernel
path_selector	Round-robin 0 or queue-length 0 for RHEL 7	Specifies the default algorithm to use in determining what path to use for the next I/O operation, round robin loops through every path in the path group, sending the same amount of I/O to each. queue-length sends the next bunch of I/O down the path with the least number of outstanding I/O requests.
path_grouping_policy	Multibus or group_by_prio(Hyperswap)	Specifies the default path grouping policy to apply to unspecified multipaths, multibus uses all paths in 1 priority group
path_checker	tur	Specifies the default method used to determine the state of the paths, tur issues a test unit ready to the device
rr_min_io OR	1	Specifies the number of I/O requests to route to a path before switching to the next path in the current path group. This setting is only for systems running kernels older that 2.6.31. Newer systems should use rr_min_io_rq, default is 1000
rr_min_io_rq(from rhel 6)	1	Specifies the number of I/O requests to route to a path before switching to the next path in the current path group, using request-based device-mapper-multipath. Default is 1

Parameter	Recommendations	Description
fast_io_fail_tmo	3	option controls how long the SCSI layer waits after a SCSI device fails before failing back the I/O. This option can be set to "off" or any number less than the dev_loss_tmo option.
dev_loss_tmo	5	option controls how long the SCSI layer waits after a SCSI device fails before marking it as failed. The default values for these options are set by the SCSI device drivers.

AIX

Table 6 on page 9 shows the AIX® parameters.

Table 6 AIX parameters

Parameter	Recommendations	Description
hdisk algorithm	shortest_queue or round_robin	Determines the methodology by which the I/O is distributed across the paths for a device, Distributes the I/O operations across multiple enabled paths. For devices that have active and passive paths, or preferred and non-preferred paths, only a subset of the paths are used for I/O operations. If a path is marked as failed or disabled, it is no longer used for sending I/O operations. The I/O operation is distributed based on path priority attribute. Paths that have a higher path priority value receive a greater share of the I/O operations. shortest_queue distributes the I/O operations across multiple enabled paths. For devices that have active and passive paths, or preferred and non-preferred paths, only a subset of the paths are used for I/O operations. This algorithm is similar to the round_robin algorithm. However, the shortest_queue algorithm distributes I/O operations based on the number of pending I/O operations on each path. The path that currently has the fewest pending I/O operations is selected for the next operation. The path priority attribute is ignored when the algorithm is set to shortest_queue
hdisk queue_depth	128	The disk queue depth limits the maximum number of commands that AIX can issue concurrently to that disk at any time. Increasing a disk queue depth might improve disk performance by increasing disk throughput (or I/O) but might also increase latency (response delay). Decreasing a disk queue depth might improve disk response time but decrease overall throughput.
hdisk max_transfer	0x100000 (1MB)	sets the limit for maximum size of the largest IO

Parameter	Recommendations	Description
hdisk max_coalesce	0x40000	sets the limit for maximum size of an individual IO that the disk driver creates by coalescing smaller adjacent requests
fcs num_cmd_elems	2048	Maximum number of requests that can be outstanding on a SCSI bus.
fcs max_xfer_size	0x200000	Maximum IO size the adapter will handle.

VMware Considerations for SQL Server

Table 7 shows the VMware parameters for SQL server.

Table 7 VMware parameters for SQL server

Description	Recommendations
VMFS	Place SQL Server data (system and user), transaction log, and backup files into separate VMDKs (if not using RDMs). The SQL Server binaries are usually installed in the OS VMDK. Separating SQL Server installation files from data and transaction logs also provides better flexibility for backup, management, and troubleshooting.
Datastore vs RDMs	Performance difference are not high enough except for high OLTP databases based on VMware testing, however RDMs are required for SQL Server AlwaysON FCI
Storage I/O Control	Consider Storage I/O Control setting for mixed VM environment
ESX HBA queue depth	Default 32-64, set to 128
ESX Disk.DiskMaxIOSize	Default 32767KB, set to 4MB
ESX PSP policy	round robin
ESX PSP IOPS limit	Default 1000, set between 1-10

Disk layout for IBM FlashSystem A9000

Figure 1 shows the disk layout for designing and mapping database volumes to IBM FlashSystem A9000 volumes, larger (2 - 4TB) luns are preferred on FlashSystem A9000. FlashSystem A9000 is globally thin provisioned and always on data reduction storage system, where only the actual data written by databases are stored on FlashSystem A9000 after pattern matching, deduplication and compression is applied. For example, if an Oracle Database writes 10 GB of table data to a 2 TB volume on FlashSystem A9000, FlashSystem A9000 effectively writes 3-4 GB of physical data (assuming 60-70% average compression ratio) to Flash drives.

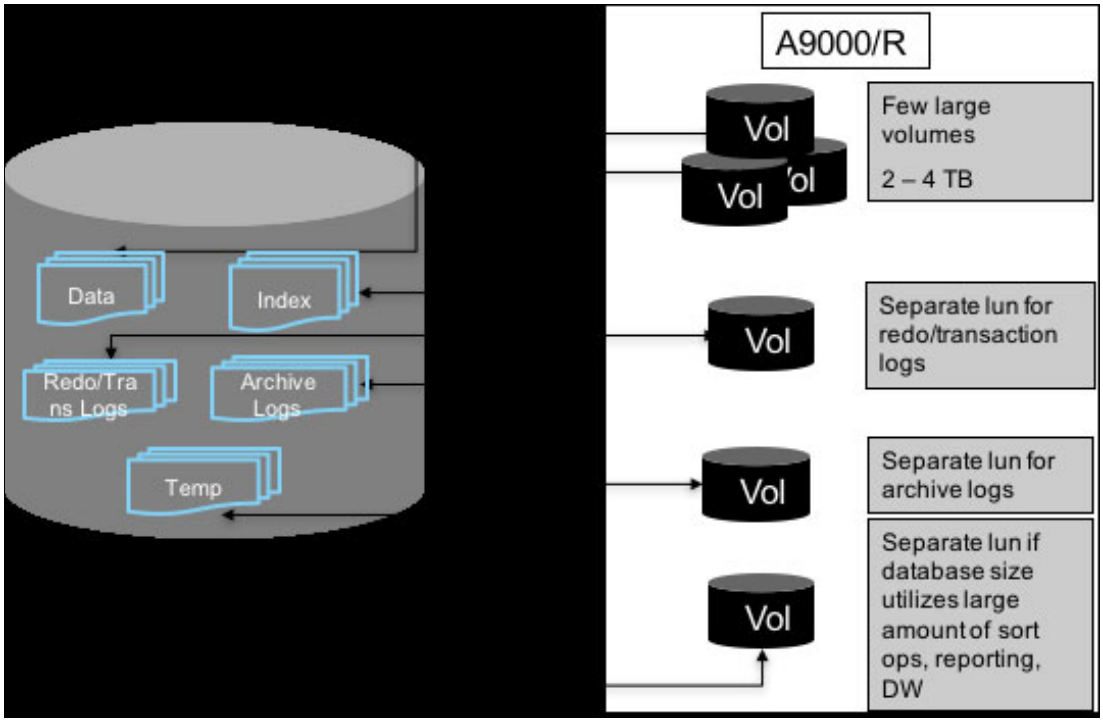

Figure 1 Disk Layout for FlashSystem A9000

Considerations for FlashSystem A9000

- Segregate data by creating separate volumes for data, logs, archive logs, backups and software install binaries
- Use fewer larger volumes for Data (2 - 4TB)
- Create dedicated volume for transaction logs/journal writes
- Create dedicated volume for archive logs
- Create dedicated volume for temp data, if database size utilizes large amount of sort operations, for example, reporting, DataWarehouse
- Additional space allocation might be required for volumes used for data compressed at database/host level
- Use QoS for non-production volumes to eliminate performance issues caused by competing workloads due to stress testing

Disk layout for FlashSystem V9000

Following illustration describes disk layout for designing and mapping database volumes to FlashSystem V9000 volumes. Balanced performance and lower latency can be achieved by using multiple vdisks for database data volumes.

Figure 2 shows the disk layout for FlashSystem V9000 databases.

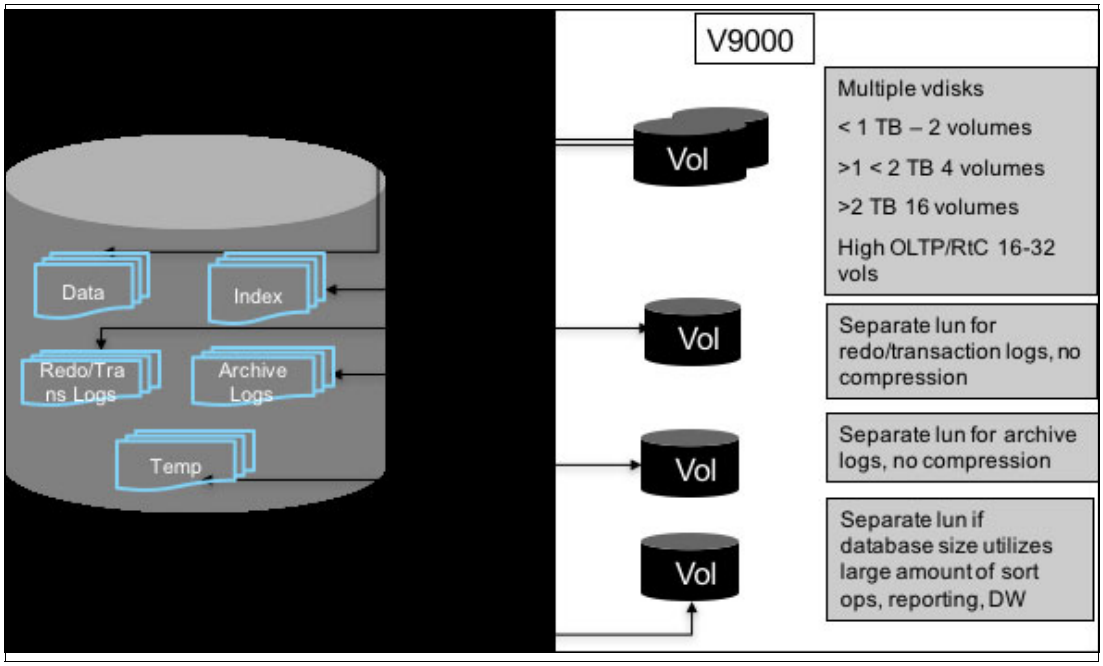

Figure 2 Disk Layout for V9000

Considerations for FlashSystem V9000

These are the considerations for FlashSystem V9000:

- Segregate data by creating separate vdisks for data, logs, archive logs, backups and software install binaries
- Use multiples vdisks for Data (16 -32) for High OLTP
- Use multiple vdisks for RtC volumes 16 - 32
- Do not compress redo or transaction Logs
- Do not compress if compression is turned on at database level
- Disable caching for volumes where write workload exceeds 1GBps
- Use QoS for non-production volumes to eliminate performance issues caused by competing workloads due to stress testing.

References

http://docs.oracle.com/cd/B19306_01/server.102/b14211/iodesign.htm#i19636
https://docs.microsoft.com/en-us/sql/relational-databases/pages-and-extents-architecture-guide
https://www.ibm.com/support/knowledgecenter/
https://technet.microsoft.com/en-us/library/cc966534.aspx

https://access.redhat.com/solutions

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an IBM Redbooks® residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks
- Follow us on Twitter:

  http://twitter.com/ibmredbooks
- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806
- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm
- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

Author

Jagadeesh Papaiah is a Corporate Solutions Architect. As a member of the IBM Worldwide FlashSystem Solutions Engineering team, he works with customers, IBM Business Partners, and IBM employees worldwide on consulting, designing, and implementing infrastructure solutions. He holds a Bachelor of Engineering Degree in Industrial and Production Engineering. He has over 23 years of experience in information management, integration architecture, infrastructure services, IT strategy & architecture, and solution design.

This project was managed by:

Jon Tate
IBM Redbooks, San Jose Center

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an IBM Redbooks residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks
- Follow us on Twitter:

  http://twitter.com/ibmredbooks
- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806
- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm
- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

REDP-5520-00

ISBN DocISBN

Printed in U.S.A.