Advanced DataTools 2017
WebCast Series

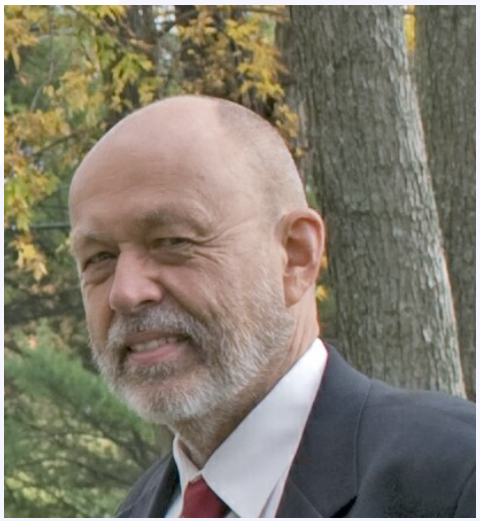
Informix Best Practices
Disks and Database Space Layout

March 30, 2017

Art S. Kagel, President & Principal Consultant
ASK Database Management

ASK Database Management

Lester Knutsen



Lester Knutsen is President of Advanced DataTools Corporation, and has been building large Data Warehouse and Business Systems using Informix Database software since 1983. Lester focuses on large database performance tuning, training and consulting. Lester is a member of the IBM Gold Consultant program and was presented with one of the Inaugural IBM Data Champion awards by IBM. Lester was one of the founders of the International Informix Users Group and the Washington Area Informix User Group.

lester@advancedatatools.com
www.advancedatatools.com
703-256-0267 x102

Art Kagel

Art Kagel is President and Principal Consultant at
ASK Database Management
an Advanced DataTools Corporation partner.

Art is a member of the IIUG Board of Directors and a recipient of the IIUG Directors Award. Art is a five time recipient of the IBM Information Champion Award. Art has over thirty years of database experience.

He has served as Manager of Database Systems and Services for a major information systems and media corporation and is a leading consultant specializing in IBM's Informix Server.

Contact Info:
art@advancedatatools.com
Cell: 732-213-5367

Advanced DataTools

Mike Walker

Mike Walker has been using Informix databases for 18 years, as a developer and as a database administrator.

Recently Mike has been developing and supporting large data warehouses for the Department of Agriculture.

Contact Info:
mike@advancedatatools.com
www.advancedatatools.com
Office: 303-838-0869
Cell: 303-909-4265

Tom Beebe

Tom is a Senior Database Consultant and has been with Advanced DataTools for over 10 years. He has been working with Informix since college with a long time fondness for open source languages. Tom is the lead consultant for Networking, Unix System Administration and Web Development needs. Currently, he is the Project Manager and lead developer on a variety of Web Development projects.

Contact Info:
tom@advancedatatools.com
www.advancedatatools.com
703-256-0267 x 106

Start With The Basics
Storage From the Bottom Up

- Media Choice
  - SSD Drives are reliable enough for databases*
  - Magnetic Drives
    - 15,000 RPM spindle speed for high access rate data
    - 10,000 RPM spindle speed drives for historical data

* Life expectancy under database conditions: about 5 years.

Start With The Basics
Storage From the Bottom Up

- Physical Configuration:
  - All disk structures used for databases MUST be RAID1 or RAID10. Period! End of discussion.

NO RAID5!

Start With The Basics
Storage From the Bottom Up

- Physical Configuration:
  - Smaller block/stripe sizes are better than larger ones.
    - 256K is the largest I would recommend
    - 128K or 64K is better

Channels

- Type of channel:
  - Copper dedicated
  - Fiber dedicated
  - Network
- Number of channels on the system
- Number of channels on the SAN/NAS
- Number of other systems sharing the SAN/NAS

Start With The Basics
Storage From the Bottom Up

- Physical Configuration:
  - Use physical arrays that are exclusively dedicated to the database
  - Avoid hierarchical storage solutions
  - Avoid virtualized storage solutions

Start With The Basics
Storage From the Bottom Up

- Interface:
  - RAW (character) devices are still best for chunks
    - Less obvious to the uninitiated as being “available”
    - Cannot build a filesystem over an open device
    - Less chance of interference from other applications (depending on SAN layouts)

Start With The Basics
Storage From the Bottom Up

- Interface:
  - Can use cooked file system files with DIRECT_IO enabled.*
    - Chunks can be extendable
    - File systems can often be expanded as needed
    - Downsides:
      - Chunks are at more risk for deletion and permanent damage
      - Harder to prevent non-database storage encroachment
      - 3-5% performance hit versus RAW

* CONCURRENT_IO on JFS2 file systems

Start With The Basics
Storage From the Bottom Up

- File System Choice
  - Avoid journaling filesystems
    - Worst journaling FS's:
      - EXT4, EXT3 (with journaling enabled), ZFS, BTFS,
    - Best journaling FS's:
      - JFS2/OpenJFS
    - On Linux: EXT2 or EXT3 with journal disabled.
  - Avoid "light weight" filesystems
    They are optimized for the speed of opening and closing many small files, otherwise no performance gain and proprietary.

Start With The Basics
Storage From the Bottom Up

- Warning!

IO Performance under virtualized environments is NEVER as good as running on bare metal!

Make sure you know your IO bandwidth requirements and test your actual throughput limits!

Storage Pool

- Large Predefined Files to be used to automate expanding dbspaces as needed.
  - Current configuration allows only a single storage pool
  - Disadvantage:
    Some applications may required multiple storage pools to allow isolating dbspaces at the storage level. Ex:
    - Transaction tables versus historical tables
    - Data tables versus indexes versus blobspace versus temp space

DBSPACES

- When you create a new dbspace consider:
  - Page size (onspaces -k)
  - Partition table extent sizing (onspaces -ef & -en)
  - Should the dbspace be expandable (automate adding chunks from the storage pool)

DBSPACES

- There are many types of dbspaces:
  - “Normal” dbspaces
  - Temporary dbspaces
  - Blobspaces
  - Unlogged Smart Blobspaces
  - Logged Smart Blobspaces
  - Physical Log Dbspace*
  - External Dbspaces

*Physical log dbspaces are supported in v12.10.xC3 and later.

CHUNKS

- When you define a new chunk, consider:
  - Should the chunk be extendable?
  - How active will the data to be stored there be?
  - What else is stored on the structure I’m considering using?
  - Where should it be placed within storage?
    - SSD
    - Fast magnetic disk
    - Slower magnetic disk
  - Should the chunk have an Informix mirror chunk?

What to put where?

- Rootdbs, logical logs, and physical log have to reside in dbspaces that are defined with the base page size (4K on AIX, Windows, & MacOS). Should be on your fastest storage!
- Placing the physical log in a PLOG dbspace allows the physical log to expand if needed to maintain transaction integrity
- Logical logs and physical logs should not be placed in the Root dbspace

What to put where?

- Place indexes on wider pages
  - IBM testing shows that nearly all indexes perform best on 16K pages
  - Allows easy isolation of index IO from data IO
  - Allows easy isolation of index cache pages from data cache pages
  - Index and data IO patterns and caching requirements are different

What to put where?

- Place tables on page sizes the minimize wasted storage
  - Minimize the slack on each page
  - Reduce or eliminate remainder pages for wide row tables
  - Minimize waste from variable length tables caused by under utilized variable length columns (ie very large LVARCHAR columns with little content on average)

```
$ waste
Usage: /home/art/bin/waste <database> [pagesize 1(K)] [pagesize 2(K)]

        Compare current table waste versus 2 other page sizes.
           Defaults to 8K & 16K.
           Only reports tables wasting more than 5% of their current page.

$ waste art
```

Table	RowSz/Pgsz	Rows Per Pg	Waste Per Pg	Per Row	Waste 8K Page	Per Row 8K	Waste 16K Page	Per Row 16K
path_test	1024(2048)	1	996	996	972	138	940	62
naf0809_rec	1270(2048)	1	750	750	524	87	1072	89
naf0809_audit	1315(2048)	1	705	705	254	42	532	44
naf0910_rec	1433(2048)	1	587	587	983	196	553	50
xserver_info	508(2048)	3	488	162	488	32	488	15
t_types	560(2048)	3	332	110	272	19	4	0
tcgslothist_new	870(2048)	2	276	138	302	33	628	34
xsql_languages	1778(2048)	1	242	242	1040	260	322	35
long_string	207(2048)	9	125	13	150	3	113	1
myptnhdr	316(2048)	6	104	17	168	6	40	0
upgrade_test	136(2048)	14	64	4	48	0	120	1
cc_pldord_conv	147(2048)	13	61	4	14	0	52	0
veh_lien_holder	127(2048)	15	59	3	46	0	116	0
sysbldiprovided	128(2048)	15	44	2	116	1	124	1
sysbldirequired	128(2048)	15	44	2	116	1	124	1
upper_test	128(2048)	15	44	2	116	1	124	1
lvtest	113(2048)	17	35	2	95	1	97	0
loadlvtest2	113(2048)	17	35	2	95	1	97	0
loadlvtest	113(2048)	17	35	2	95	1	97	0

```
$ new_waste_all.ksh adtc_monitoring

Table: client_master Variable Avg Row Len: 31 Num Rows: 5  Max Len: 269

      Page Size         Row Sz        Rows/Pg       Waste/Pg      Waste/Row
           2048             31             55            202              3
           4096             31            111            407              3
           6144             31            166            642              3
           8192             31            222            816              3
          10240             31            255           1801              7
          12288             31            255           3849             15
          14336             31            255           5897             23
          16384             31            255           7945             31


Table: contacts Variable Avg Row Len: 151 Num Rows: 4  Max Len: 164

      Page Size         Row Sz        Rows/Pg       Waste/Pg      Waste/Row
           2048            151             11            202             16
           4096            151             23            407             17
           6144            151             36            612             17
           8192            151             48            816             16
          10240            151             60           1021             17
          12288            151             72           1226             17
          14336            151             84           1431             16
          16384            151             96           1672             17


Table: checkpoint_stats Fixed Row Len: 270 Num Rows: 65

      Page Size         Row Sz        Rows/Pg       Waste/Pg      Waste/Row
           2048            270              7            120             17
           4096            270             14            264             18
           6144            270             22            136              6
           8192            270             30              8              0
          10240            270             37            152              4
          12288            270             45             24              0
          14336            270             52            168              3
          16384            270             60             40              0
```

Partitioning

- Why partition?
  - Avoid/alleviate the 16 million page partition limit *
  - Divide and conquer
    - Allow parallel query
    - Allow partition elimination
    - More accurate data distributions for larger tables
  - Storage placement of different classes of table content
    - Current rows on faster storage
    - Historical rows on slower/cheaper storage

* The maximum number of data pages in a partition is 2^24 -1 = 16,777,215 pages because the last 8 bits of the ROWID is the slot number of the row on the page.

Partitioning Schemes

- ROUND ROBIN
  - Simple
  - Fastest insert speed for insert-heavy tables with many insert client sessions
  - No partition elimination
  - Adding or dropping partitions may require rewriting the entire table
  - For tables only (not for index use)

Partitioning Schemes

- BY EXPRESSION
  - One or more BOOLEAN expressions defining what rows are to be placed in each partition
  - Zero or One “REMAINDER” partition holding rows that do not match any partition expression
  - Expressions are evaluated sequentially in the order defined.
  - Long lists of expressions can be expensive to process at run time

Partitioning Schemes

- BY LIST (column-expression)
  - PARTITION <name> VALUES (list), ...
  - PARTITION <name> IS NULL - optional
  - PARTITION <name> REMAINDER - optional
- Internally hashed so it can be faster to process than BY EXPRESSION partitioning when the expressions all refer to the same column(s)

Partitioning Scheme

- FRAGMENT BY RANGE(column)
  - Can use a numeric, DATE, or DATETIME column
  - INTERVAL(<N units expression>)
    - STORE IN (<dbspace1>, <dbspace2>, …) - optional
    - STORE IN (<function returning a dbspace name>) - optional
  - PARTITION <name> VALUES **<** <number> IN <dbspace>

Partitioning Scheme

- Rolling window partition management:

  Range partition schemes can also automatically add and drop partitions to maintain a predefined number or range of active partitions:
  - ROLLING (<# partitions> FRAGMENTS)
    - DETACH
    - DISCARD
  - ROLLING (<# partitions> FRAGMENTS)
    LIMIT TO <N> <Units> (Units: KB, MB, GB, TB)
    - DETACH
      - INTERVAL ONLY
      - INTERVAL FIRST
      - ANY
    - DISCARD
      - INTERVAL ONLY
      - INTERVAL FIRST
      - ANY

Partitioning Schemes

- Instead of partitioning an index on a range or expression of key values, you can define the index as a FOREST OF TREES index:
  - CREATE INDEX <indexname>
    ON <tablename>(<key list>)
    HASH ON (<leading key list>)
    WITH <num> BUCKETS
- FOT indexes create <num> B+Trees each with fewer levels
- Contention for the root node of each sub-tree is expected to be less than for the root of a single B+Tree index would be assuming client access to the hashed part of the key is well distributed across the sub-trees.

Extents

- Prior to version 11.70 extent information for a partition had to fit on the single partition header (or tablespace tablespace) page limiting a partition to about 360 extents (on 2K pages).
- For v11.70 and later the partition header page can be automatically extended by additional pages making the new limit 32765 extents which with extent doubling and the 2^24 pages limit on partitions makes the maximum number of extents effectively unlimited.

Extents

- However, we still have to be concerned when a partition has many extents if many of those extents contain actively accessed data
  - In those cases normal operations will be stressing the IO subsystems.
  - On magnetic disk, head movement latency will exacerbate the issue.
- Less important on SSD drives, however, many small active extents can reduce the usable life of your SSD memory cells which may ultimately cause data loss.

Extents

- Ideal extent sizing:
  - The initial and next extent sizes should cover the active working set of data in a very small number of large extents.
  - If you are defragmenting an existing table with many extents, set the EXTENT SIZE to hold a large subset of the existing data, perhaps even all of it.
  - Set the NEXT SIZE as you would for a new table, to hold the normal working set or perhaps half of the working set if that set is large.

Let’s Get Back to My Favorite Subject!

RAID5 bashing!

??? NO RAID5 ???

- Why is RAID5 more popular than RAID10 among storage administrators?

1. It requires fewer drives to present the same storage capacity.
2. Storage sales people can present a less expensive proposal to meet the required storage volume. They make a slanted case to make a sale!
3. Most people have never studied the issue themselves and so trust that RAID5 is good.

??? NO RAID5 ???

- What's wrong with RAID5?
  - Least important issue:
    The "RAID5 write penalty"
    - RAID5 write performance is 30-60% slower than RAID0 for an array with the same capacity. Even more so when compared to RAID10.
  - Most important isue:
    RAID5 is not safe. Your data is at risk!

??? NO RAID5 ???

- What's wrong with RAID5?
  - Recovery from a failed RAID5 drive takes MUCH longer to recover than a failed RAID10 drive. All of the drives in the array have to be read and either new checksums or reconstructed data has to be calculated before writing.
  - Performance of the array as a whole degrades while a RAID10 drive is offline by a percentage equal to: 100% / <num drive pairs> (ex: 20% for 5 pairs, 10% for 10)
  - Performance degrades by over 80% while a RAID5 replacement drive is being repopulated using checksum data.

??? NO RAID5 ???

- What's wrong with RAID5?
  - If any other drive in the array is lost before the repopulation of the replacement drive completes all of the data on the entire array is a total loss. It is impossible to recover a complete, coherent, data set from the remaining drives.
  - Most drive failures are caused by hardware and firmware failures which are likely to have affected multiple drives in the array. A second failure is more likely than not. The more drives in the array, the more likely catastrophic data loss is.

??? NO RAID5 ???

- What's wrong with RAID5?
  - Magnetic drives suffer partial media failure as they near their end-of-life.*
  - They will eventually run out of remap sectors and the damaged sectors will no longer be correctable. Since RAID5 does not validate its checksums on read, if an undamaged drive fails the RAID5 firmware will rebuild it from data and checksums on the damaged drive creating incorrect data on two drives, one of which is new.

* About 4-5 years.

??? NO RAID5 ???

- What's wrong with RAID5?
  - SSD drive memory cells have a finite lifetime counted in write refresh cycles.*
  - The built-in firmware extends the lifetime of the "drive" as a whole by never rewriting a cell in-place. Rewrites are accomplished by writing the updated data to an unused sector (some are reserved for this purpose outside the reported capacity) is written to and the file's meta data is remapped and itself rewritten (triggering a cascade of writes).
  - Databases write far more frequently than file based applications and they write relatively small "pages" compared to the cell sizes in most SSD drives which causes a cell's data to be rewritten far more often than is the case for the test applications that are used to determine the MTBF and lifetime for these drives.

* Equivalent to about 4-5 years for database type applications.

!!! Yes RAID10 !!!

- How is RAID10 better:
  - Every drive has a mirror drive with a complete copy of the data.
  - RAID10 can read different sectors from both sides of each mirrored pair in parallel, nearly doubling read performance over a simple RAID0 array during peak loads.
  - Writes to RAID10 arrays are 100% faster than writes to RAID5 and the increase is sustainable over time.
    - RAID5 vendors counter with more cache memory on the SAN which is far more expensive than the extra drives needed to implement RAID10 instead to get the performance boost!
  - Everyone in the industry not selling arrays agrees that RAID10 is safer.

!!! Yes RAID10 !!!

- How is RAID10 better:
  - RAID10 failed drive recovery is just a copy from one drive to another. Faster than RAID5 recovery.
  - The performance impact during recovery of a RAID10 array impacts only IO from/to that one drive pair. All other drive pairs in the array continue to perform at full speed.
  - Performance degradation while a failed drive is offline only affects peak performance of that one drive pair out of the entire array. Normal IO volume performance is unaffected.

!!! Yes RAID10 !!!

- How is RAID10 better:
  - A RAID10 array can survive the loss of 50% of its drives without data loss as long as one drive from each mirrored pair is still online.
  - Pairing each drive with another from a different manufacturer’s build lot minimizes the possibility that faulty hardware or firmware will cause data loss. (Not feasible with RAID5 as drives from many different lots would be required. RAID10 only needs to access two different drive lots for safety.)

!!! Yes RAID10 !!!

- How is RAID10 better:
  - Partial media failure does not pose the same risk of data loss under RAID10 as it does under RAID5:
    - If a sector is completely unrecoverable, the drive will be marked offline and the data reread from the mirror where this sector is likely fine. This would however be a red flag to replace both drives in the pair (and all other drives of the same age) as quickly as possible.
    - Since RAID10 only ever copies data from one drive to another during recovery, it is far less likely that unrecoverable garbage will be written to a good drive.

!!! Yes RAID10 !!!

- How is RAID10 better:
  - SSD drives:
    - Because checksums do not have to be calculated and written to other drives the number of write cycles per drive in a RAID10 array is significantly lower* than in a RAID5 array resulting in a more consistent and longer lifespan.

* about half

Need More?

- Testing at CERN after they experienced data loss on RAID5 arrays determined:
  - Most drive failures (80%) are caused by hardware and firmware failure (another 10% from wrong firmware version).
  - Partial media failure accounts for much of the rest of the data loss they experienced on both magnetic and SSD drives as they aged
  - They experienced cosmic ray damage flipping bits, equally on both magnetic and SSD type drives

Need Even More?

- Drives today are commodity priced.
  Is the risk of data loss worth the relatively small savings?
- The failure rates and expected lifespans for “premium” drives are identical to commodity retail drives.
  So, it doesn’t help that you are spending >$1000 per drive
- According to a storage industry study – failure of a second drive is 4X more likely than the single drive failure rate would predict!
- Atomic writes across multiple drives in a RAID5 array are not guaranteed!

Need Even More?

- Larger drives take longer to rebuild increasing the likelihood of losing a second drive.
- A recent study concluded that drives over 1TB are statistically likely to suffer from unrecoverable multiple bit dropouts.

  The number of bits on the drive exceeds the bit failure rate!
- The error rates as observed by the CERN study on silent corruption, are far higher than the official rate of one in every 10^16 bits.

  The observed error rate was about one in 10^7 bits or 1 out of about 1 in every 1,000,000 bits (~125,000 bytes).

!!! NO RAID5 !!!

Informix Best Practices
Disks and Database Space Layout

Art S. Kagel

www.askdbmgt.com

art@askdbmgt.com
art.kagel@gmail.com
art@advancedatatools.com

ASK Database Management

Next Webcast
Informix Best Practices

- Backup, Recovery, and High Availability Disaster Recovery by Lester Knutsen
  - Thursday, April 20, 2017 at 2:00pm EST
- Informix Configuration, ONCONFIG part 2 by Lester Knutsen
  - Thursday, May 18, 2017 at 2:00pm EST
- Informix Connection Manager by Thomas Beebe
  - Thursday, June 29, 2017 at 2:00pm EST
- Informix Auditing by Mike Walker
  - Thursday, July 27, 2017 at 2:00pm EST

Please register for each webcast here at:
http://advancedatatools.com/Informix/NextWebcast.html

Informix Resources - IIUG

- The International Informix User Group
  - http://www.iiug.org
  - Membership is FREE
- IIUG 2017 – The Premier Informix Event
  April 23 – 27, 2017
  - http://www.iiug2017.org

Informix Resources from IBM

- Informix Documentation
  http://www.ibm.com/support/knowledgecenter/SSGU8G_12.1.0/com.ibm.welcome.doc/welcome.htm
- Compare the Informix Version 12 editions by Carlton Doe, IBM
  http://www.ibm.com/developerworks/data/library/techarticle/dm-0801doe/
- The Informix and IoT Roadshows by Carlton Doe, IBM
  https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/North%20America%20Informix%20Events

Informix Training in 2017

- April 10-13, 2017 – 2 Seats still available!
  - **Informix for Database Administrators**
- July 10-13, 2017
  - **Advanced Informix Performance Tuning**
- September 18-21, 2017
  - **Informix for Database Administrators**
- All courses can be taken online on the web from your desk or at our training center in Virginia.
- We guarantee to *NEVER* cancel a course and will teach a course as long as one student is registered!

Questions?

Send follow-up questions to
lester@advancedatatools.com

Advanced DataTools

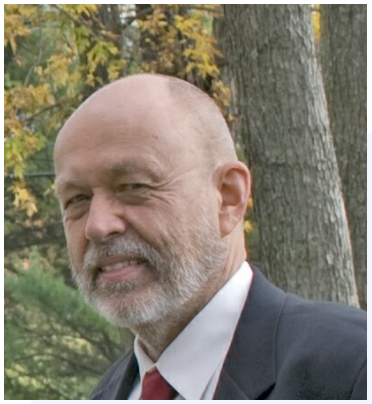
Informix Support and Training from the Informix Champions!

Advanced DataTools is an Advanced Level IBM Informix Data Management Partner, and has been an authorized Informix partner since 1993. We have a long-term relationship with IBM, we have priority access to high-level support staff, technical information, and Beta programs. Our team has been working with Informix since its inception, and includes 8 Senior Informix Database Consultants, 4 IBM Champions, 2 IIUG Director's Award winners, and an IBM Gold Consultant. We have Informix specialists Lester Knutsen and Art Kagel available to support your Informix performance tuning and monitoring requirements!

- ***Informix Remote DBA Support Monitoring***
- ***Informix Performance Tuning***
- ***Informix Training***
- ***Informix Consulting***
- ***Informix Development***

Free Informix Performance Tuning Webcast replays at:
http://advancedatatools.com/Informix/Webcasts.html

Call: (800) 807-6732 x101 or Email: info@advancedatatools.com
Web: http://www.advancedatatools.com



Thank You

Lester Knutsen

Advanced DataTools Corporation

lester@advancedatatools.com

For more information:

http://www.advancedatatools.com


IIUC 2017
The Premier
Informix Event
Raleigh, NC,
April 23rd - 27th
IBM