

SQLskills Immersion Training

Custom IE1 + IE2 *for* Bloomberg

Kimberly's Whiteboard Drawings and Annotations

Kimberly L. Tripp
Kimberly@SQLskills.com



NOTE: Includes whiteboard drawings from prior IE courses
+ drawings done the week of April 1, 2013
New York, NY - Bloomberg

SQLskills Immersion Event

IE1: Internals and Performance

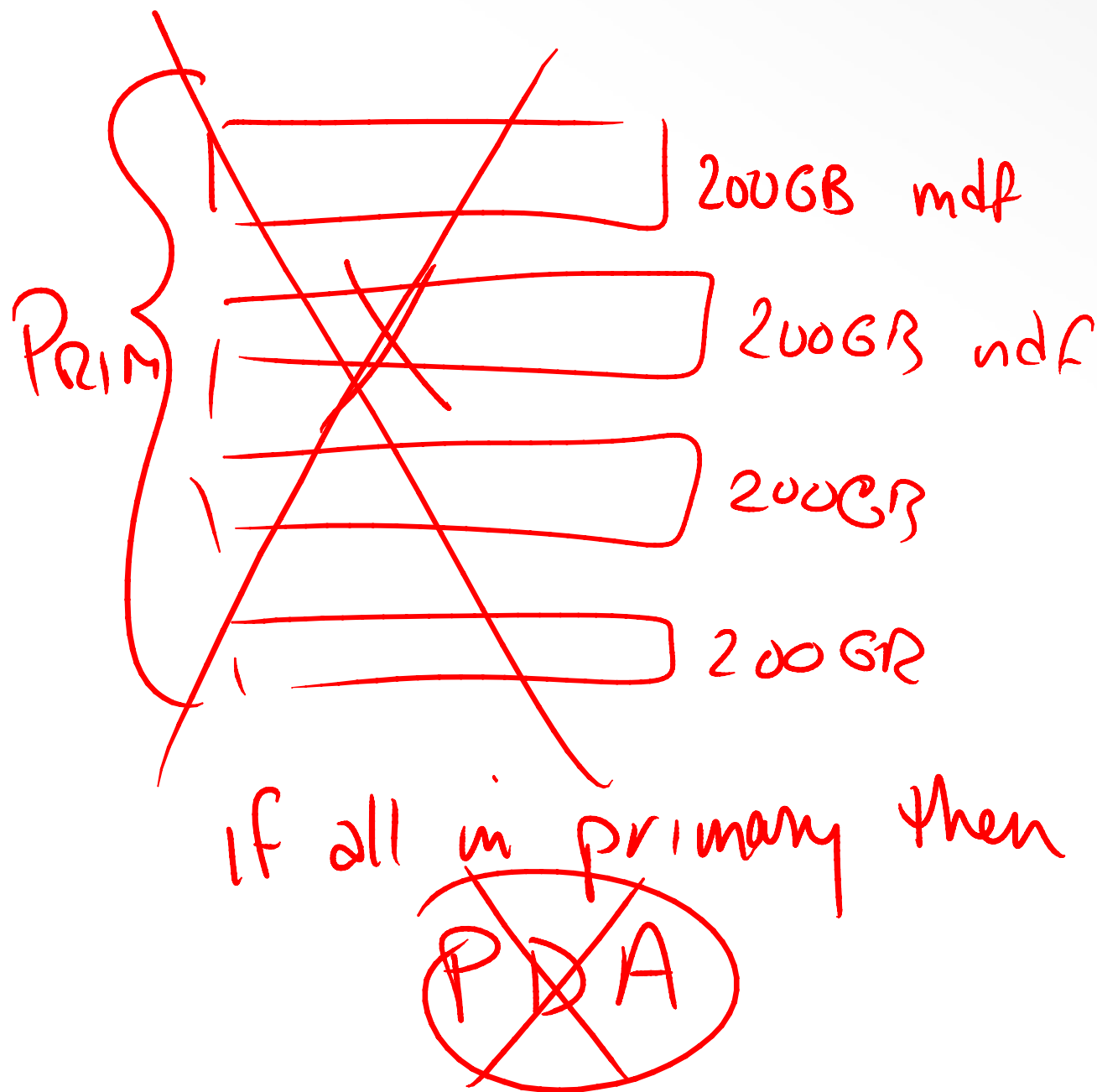
General Internals

Kimberly L. Tripp

Kimberly@SQLskills.com



NOTE: During Module 1 (Index Internals), we discussed *general database internals*; this first section is from these drawings.



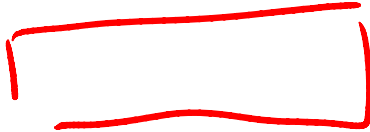
One of the problems in allowing files to be added to the primary filegroup is that it makes it more vulnerable to downtime. SQL Server's Partial Database Availability feature ONLY works when both the primary filegroup AND the log are available. If the primary filegroup has multiple files and one of those files fails – the entire database will be offline until restored.

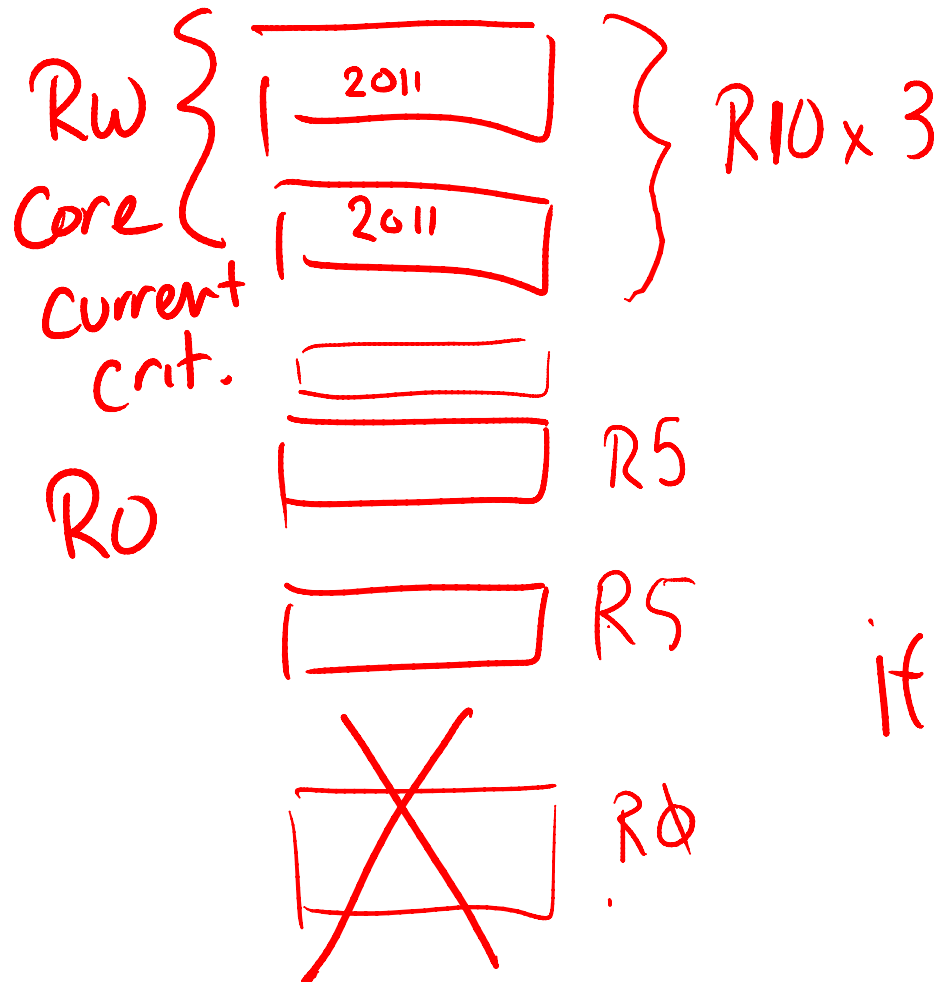
For VLDBs, you want to isolate the primary filegroup and dedicate it only to system tables. Then, place it on redundant disks.

Secondary, non-primary filegroups are where you'll store data.

(see the next drawing)

If VLDB

 mdf just SYS RID x 3



PDA
if file damaged
only that fg
offline

By dedicating only the system objects to the primary filegroup (and limiting this to only one, relatively-small mdf, you reduce the potential for downtime.

If a file does become damaged then the database can remain online and available (PDA). And, then, using Online Piecemeal Restores, you can keep your database online for most of the recovery process.

NOTE: Remember in my USB demo... there is one caveat. To take a file offline (in order to begin the restore process), you must take the filegroup of which it's a member offline. This is NOT an online operation (because SQL Server does not track transactions at the filegroup level). So, all connections to the database are terminated when you take a filegroup offline.



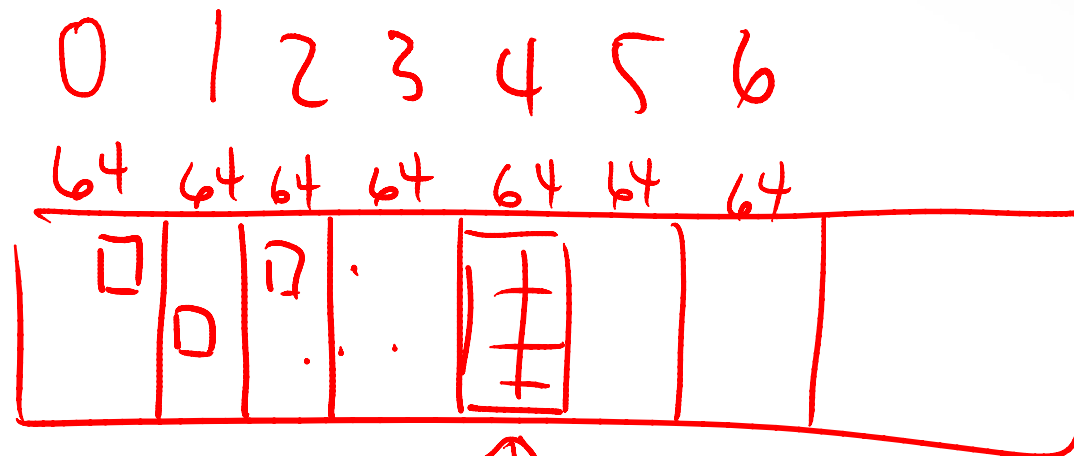
allocations?

After eight 8KB pages have been allocated to a table, SQL Server does extent-based allocations. These allocations are done file by file. Or, in other words, SQL Server round-robins extent allocations from the files in the filegroup.

This helps to reduce IO contention and also contention on the system allocation pages (GAM pages are used for extent allocations) themselves.

To the left – the first 4 pages in a file are:
Header (0)
PFS: Page Free Space (1)
GAM: Global Allocation Map (2)
SGAM: Shared Global Allocation Map (3)

IAM Pages

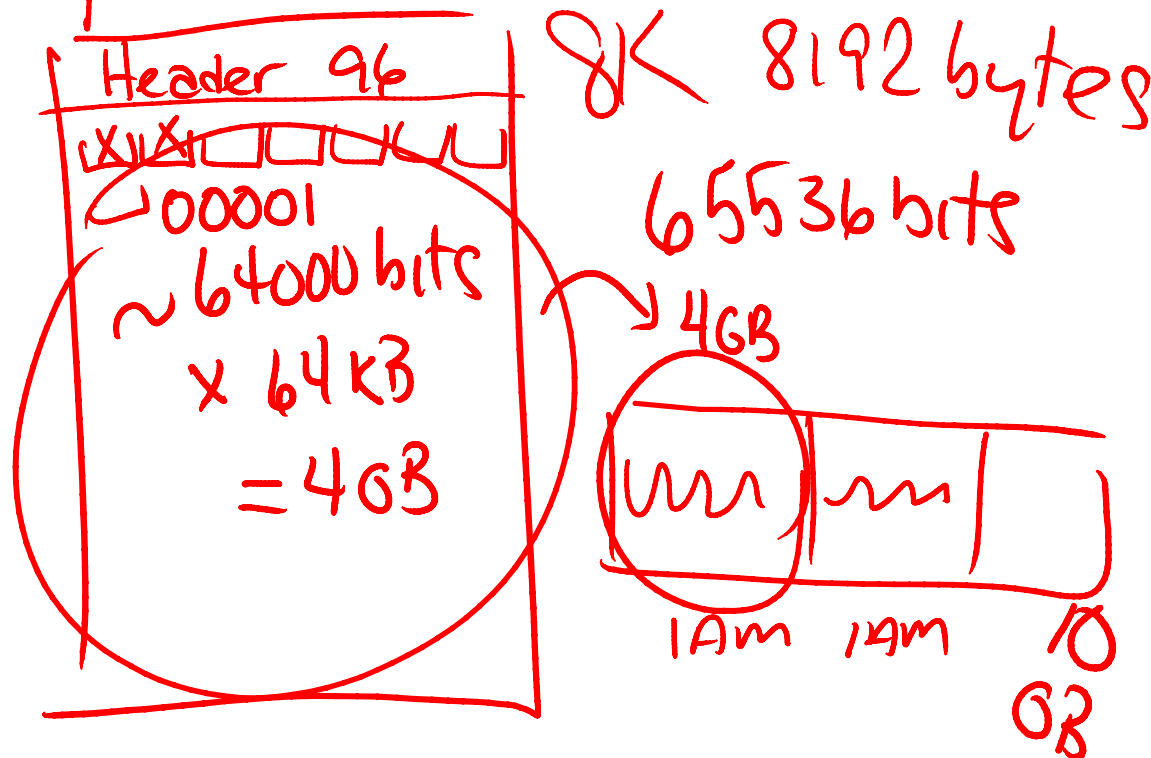


While GAM pages track extent usage within the database, IAM pages track a table's allocated extents within a GAM interval (a 4GB chunk of the file). GAM page and IAM pages can track 4GB because they use a bitmap structure on the 8KB page.

8KB = 8192 bytes = 65536 bits

With roughly 64000 bits each tracking an extent (64KB), you can see how they can track a 4GB chunk of the file.

IAM
Index Alloc Map

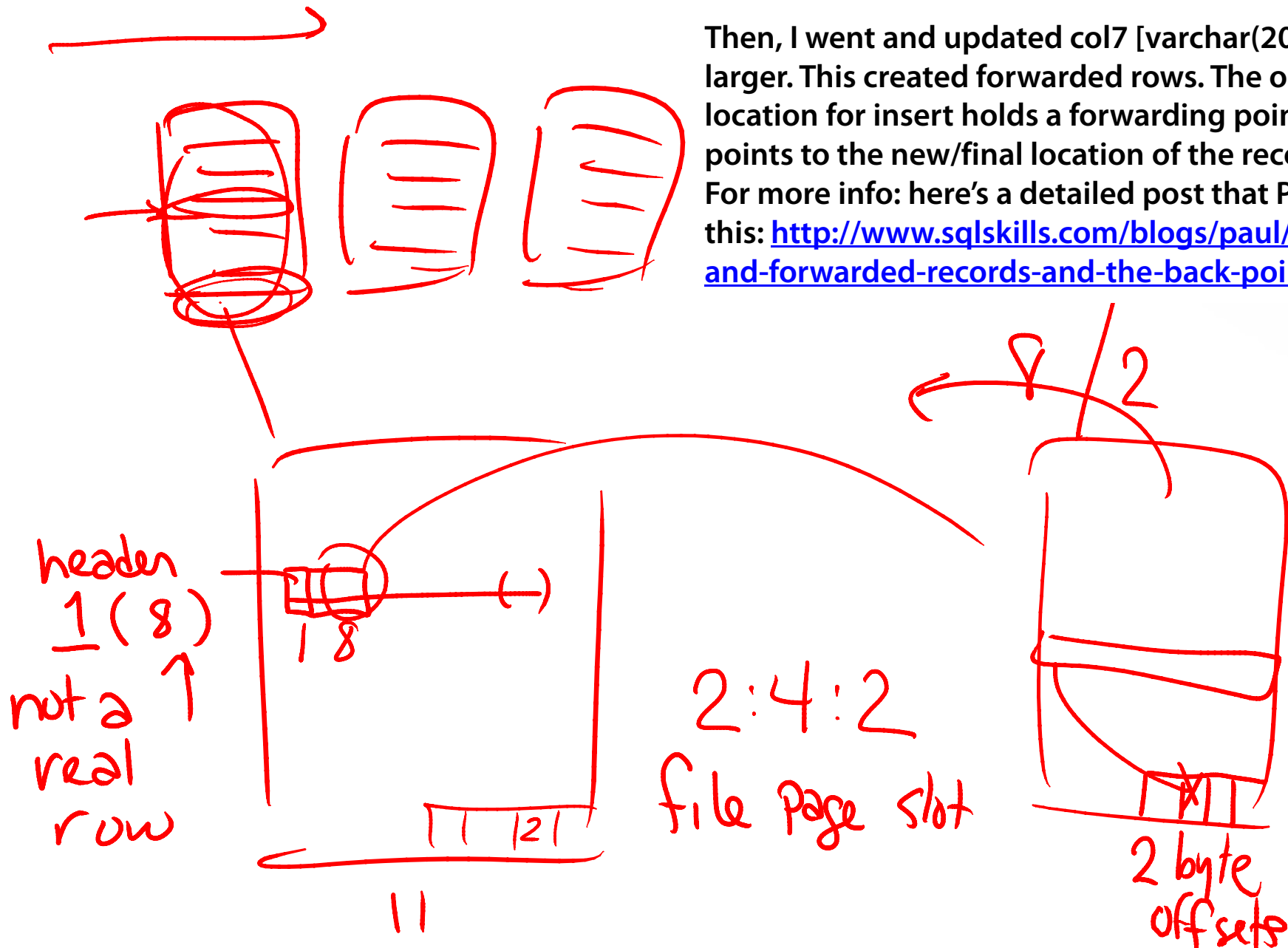


From the Heap DEMO

In my heap demo, I first added rows (as fast as possible) in the table. All of the rows were fixed width and column 7 had only 11 bytes per row. SQL Server filled the pages as fast as they could...

Then, I went and updated col7 [varchar(200)] to be larger. This created forwarded rows. The original location for insert holds a forwarding pointer and that points to the new/final location of the record.

For more info: here's a detailed post that Paul did on this: <http://www.sqlskills.com/blogs/paul/forwarding-and-forwarded-records-and-the-back-pointer-size/>

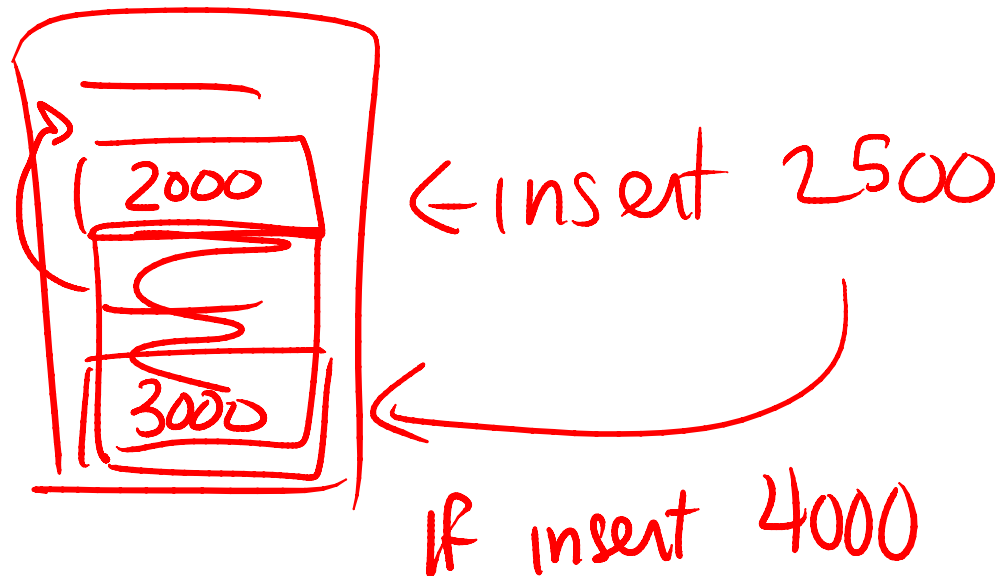
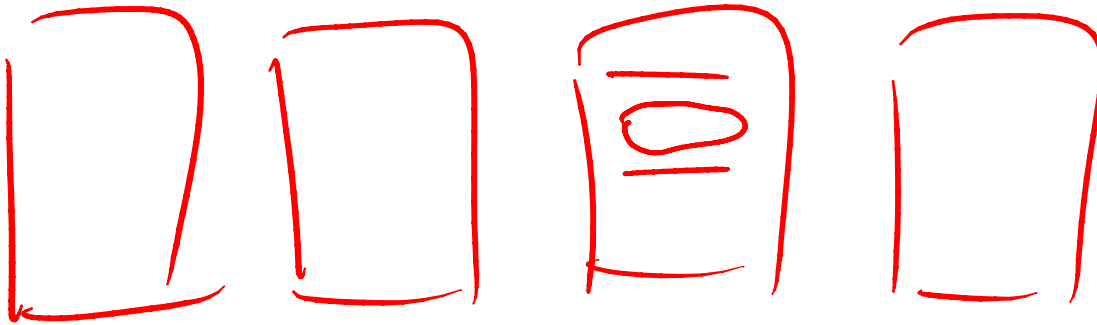


Inserts on a page (1 of 2)

SQL Server does not rearrange the records on a page – unless it needs to.

If there are two records on a page but there's a gap large enough to fit the record being inserted, then SQL Server will put the record wherever it fits.

On a heap, it will receive the next slot available. On an indexed page, the slots will be ordered by the index key.

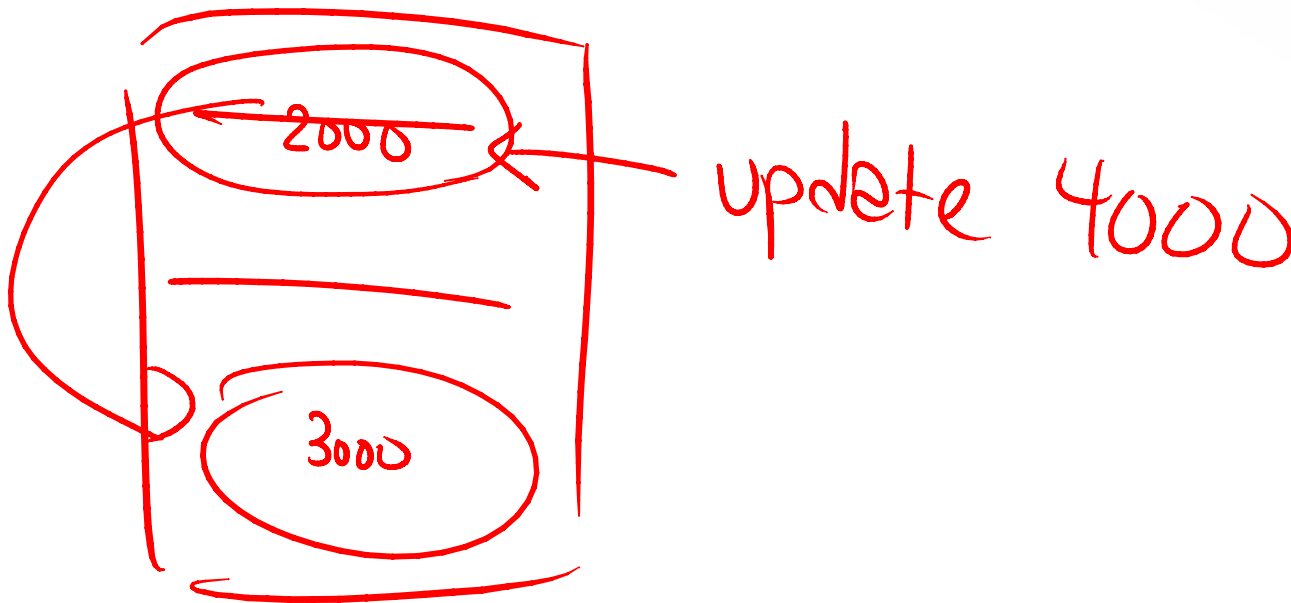


Inserts on a page (2 of 2)

SQL Server does not rearrange the records on a page – unless it needs to.

The same holds for an update. If the first record is being updated and there's not enough room without rearranging the page then SQL Server will rearrange.

If the record moves then the slot doesn't change but the offset will get updated.



Record Structures (1 of 2)

In_row fixed-width
All structures if ≤ 8060
note! never span pages
always be cont

Overflow limited VC Cds
Short LOBs
if row size > 8060 then
1+ cds w/overflow

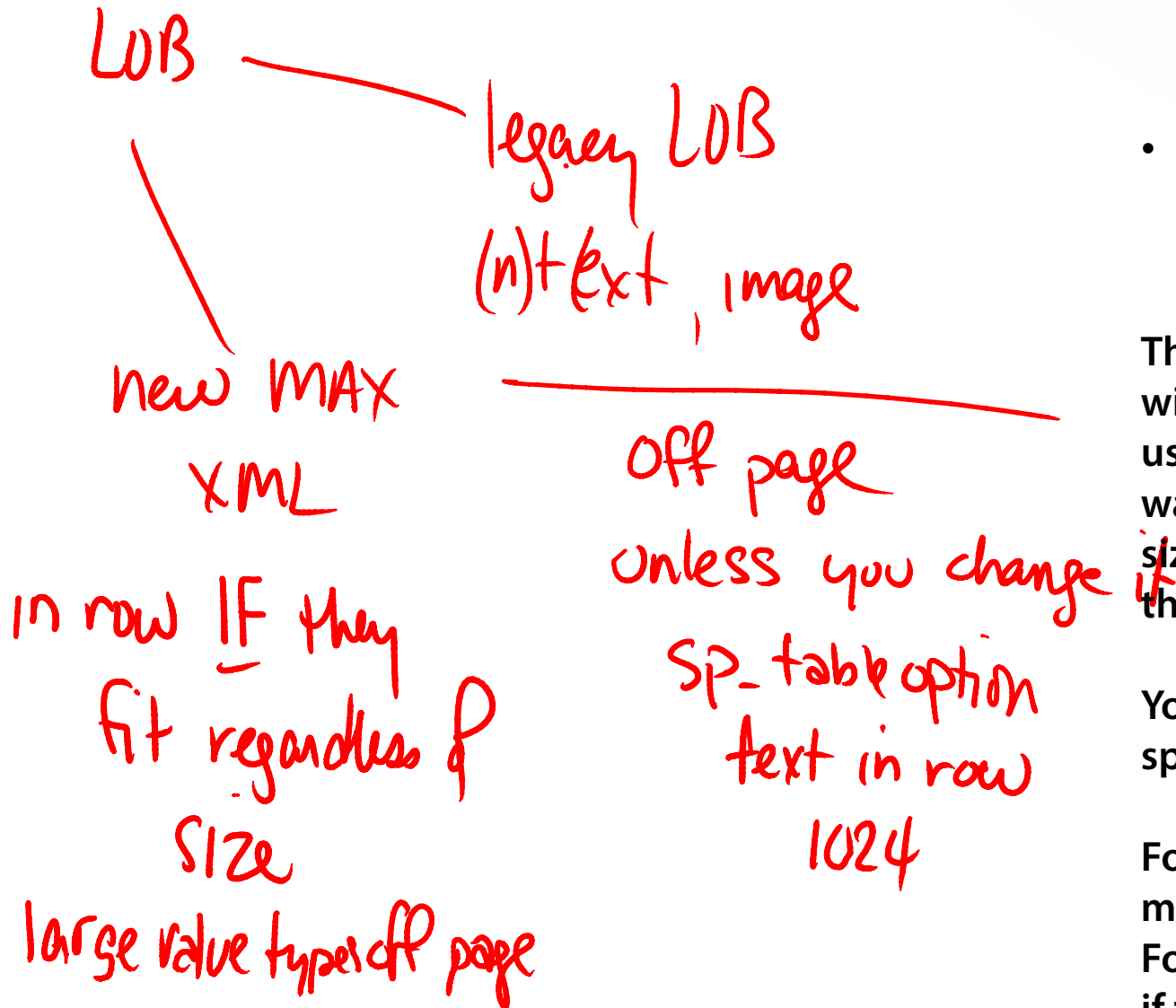
Every table/index can have up to 3 data structures associated with it:

* In_row – every row has an in_row_data structure. This cannot exceed 8060 bytes and must exist as a single, contiguous structure on – at most – one page. If a row has limited variable columns [varchar(n), nvarchar(n) or varbinary(n)] or LOB types [new LOB types: varchar(max), nvarchar(max) or varbinary(max); Legacy LOB types: text, ntext, image; XML] then you *might* have additional structures.

* Overflow – Overflow is only for limited variable columns (often called SLOBs for short LOBs). A row will ONLY overflow to the overflow structure IF the row is over 8060 bytes AND only these limited LOBs will overflow to the overflow structure. Fixed-width columns will NOT overflow. An overflowed column will always overflow as a unit and the column(s) that overflows is not easily predictable.

* LOB – explained on next page.

Record Structures (2 of 2)



* LOB – There are 8 LOB types in SQL Server:

- New LOB types
 - varchar(max)
 - nvarchar(max)
 - varbinary(max)
 - XML
 - User-defined SQLCLR types
- Legacy LOB types
 - text
 - ntext
 - Image

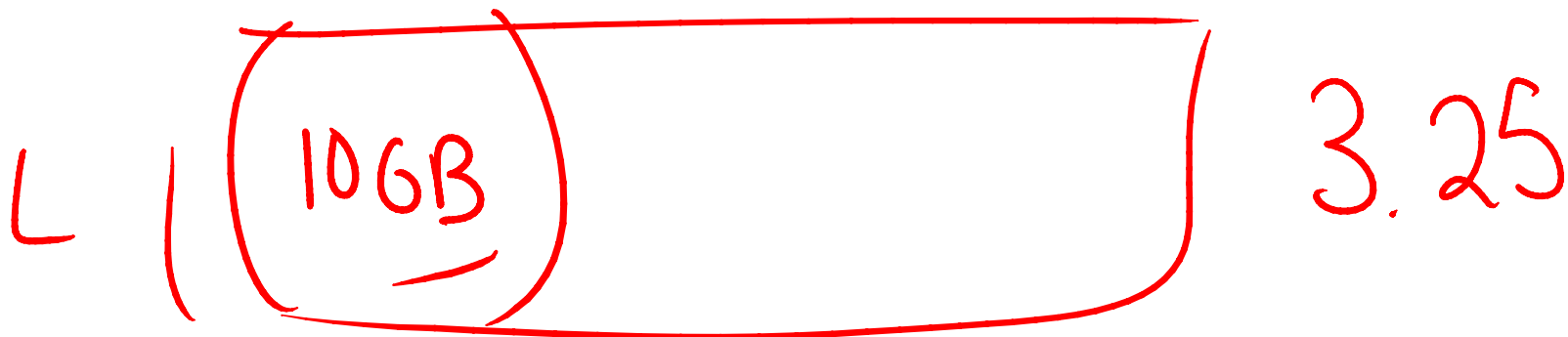
The initial location of the LOB (on insert) will depend on the LOB data type you've used. The default behavior to legacy LOB was to push them off row – regardless of size. For the new LOB types they will put them in_row only if they fit.

You can control these options by using `sp_tableoption` for every table change.

For legacy LOB you can control the maximum byte length for in_row LOBs. For new LOB types they either go in_row if they fit (default) OR you can say they go off-page no matter what.

The cost of creating a clustered index

To create a clustered index, SQL Server needs to copy the data (in the filegroup that is the destination) as well as use sort space to order the data (also in the destination). For a 10GB table this requires 10GB for the copy and ~25% for the sort space. Then, there's logging – in a fully logged database (the default), this is essentially a "size of data" operation. So, for a 10GB table, this is a 10GB operation in the log. The total size for a create or rebuild in a fully logged database is 3.25x the table size.

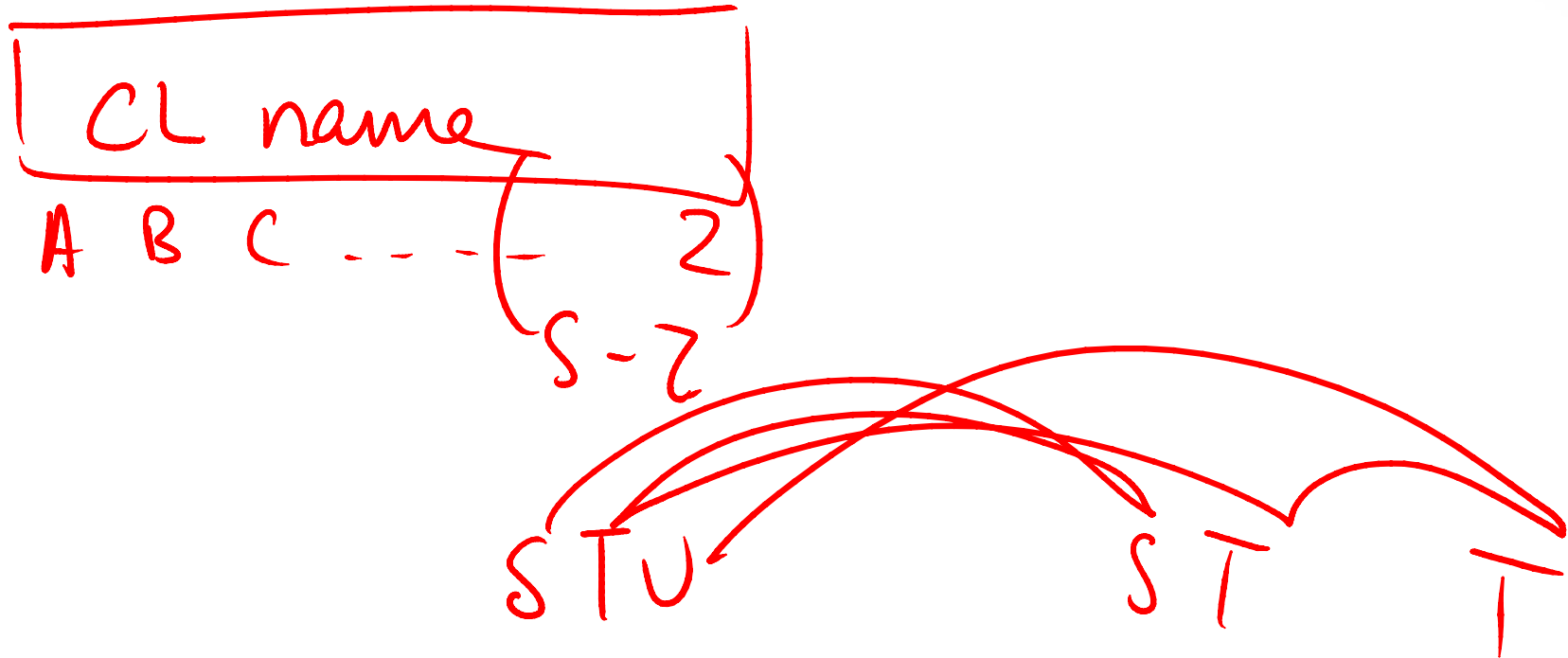


Clustering for ranges (looks good on paper)

If I show you a clustered structure and talk ONLY about querying for ranges, I can convince you that you should create a clustered index on something like Lastname, Firstname, Middleinitial

And, while this looks good on paper, it doesn't last very long. As soon as inserts/updates occur, this table will become very fragmented. This fragmentation will offset (and begin to overshadow) the apparent benefits of having chosen that order.

And, I haven't even mentioned the expense of a wide clustering key in the nonclustered indexes.



SQLskills Immersion Event

IE1: Internals and Performance

Module 7: Index Internals

Kimberly L. Tripp
Kimberly@SQLskills.com



CL must unique

LN

666

Smith
Smith 1

SSN

Jone
Jones 1

The Clustering key **MUST** be unique

The key reason that the clustering key must be unique is because of the lookup that must be performed when looking up a corresponding row from a nonclustered index request.

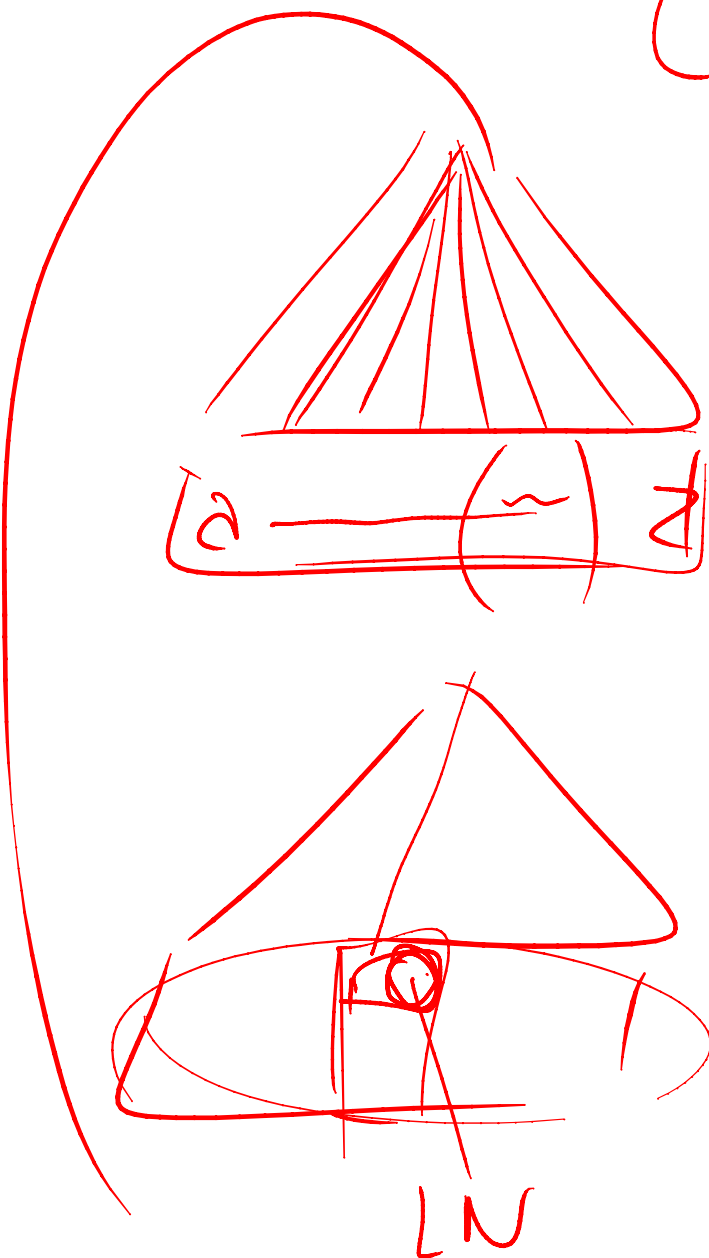
The uniquifiers are unique per duplicate key value

The FIRST Smith that is inserted is unique – so no uniquifier

The second Smith goes in and it becomes Smith 1

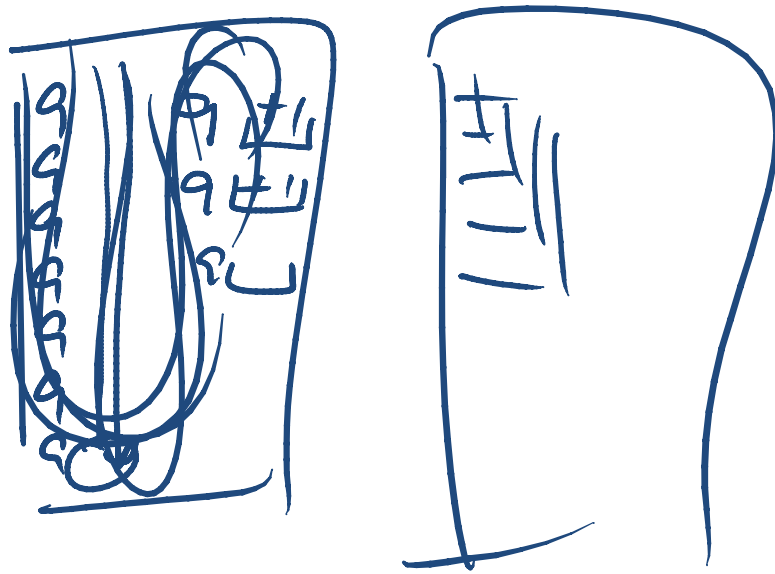
The FIRST Jones that is inserted is unique – so no uniquifier

The second Jones goes in and it becomes Jones 1



CL DATE

Int 2.147 bill



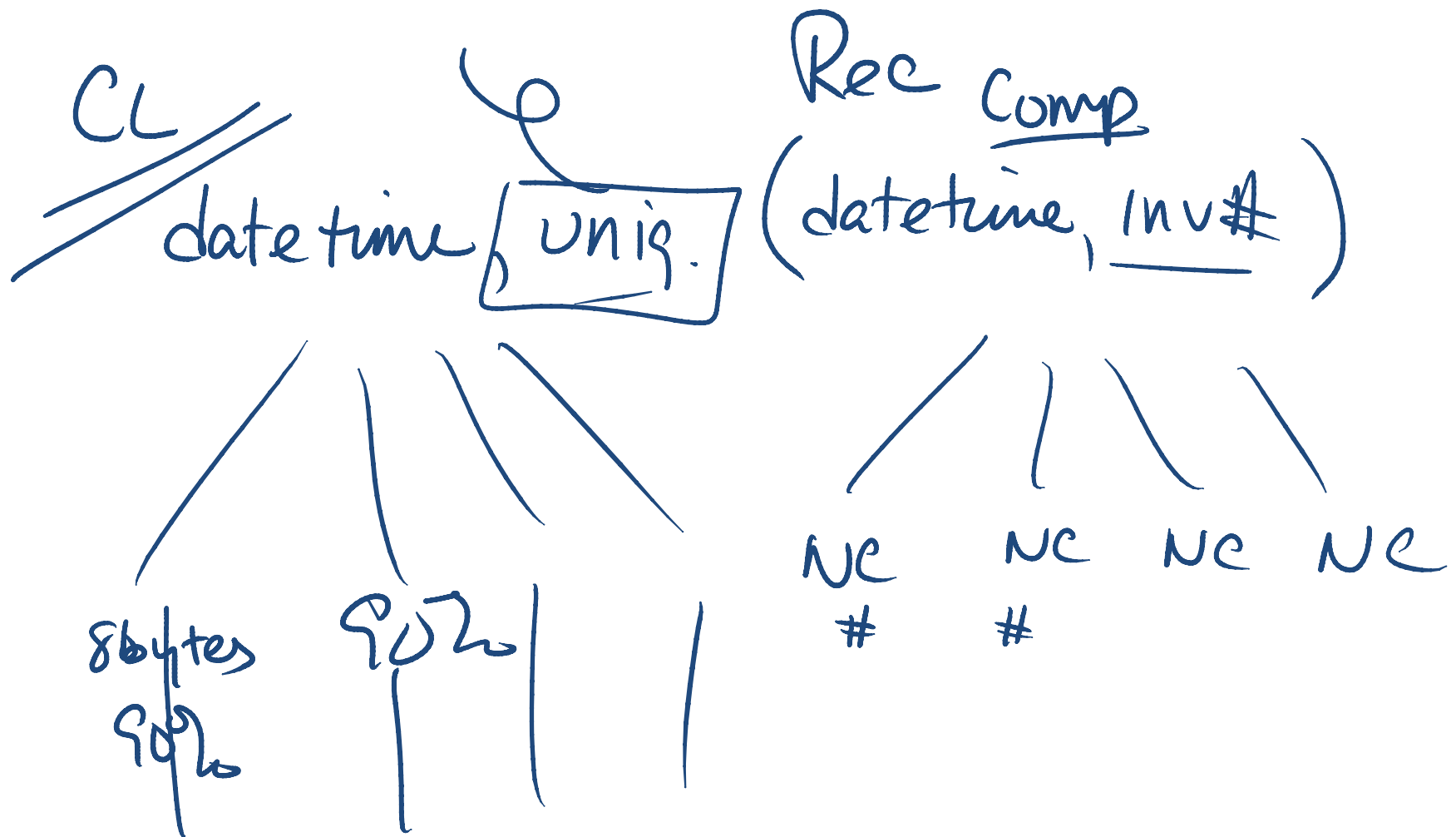
$\varepsilon_{r \text{ rov}}$

666

A clustering key should be unique by definition. However, if the CL key is NOT unique then SQL Server will uniquify it. And, this process of uniquification can be expensive because the next uniquifier value is unknown (they must scan the last page of values in that duplicate range).

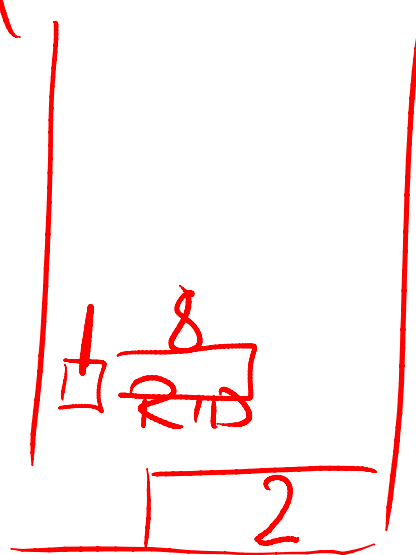
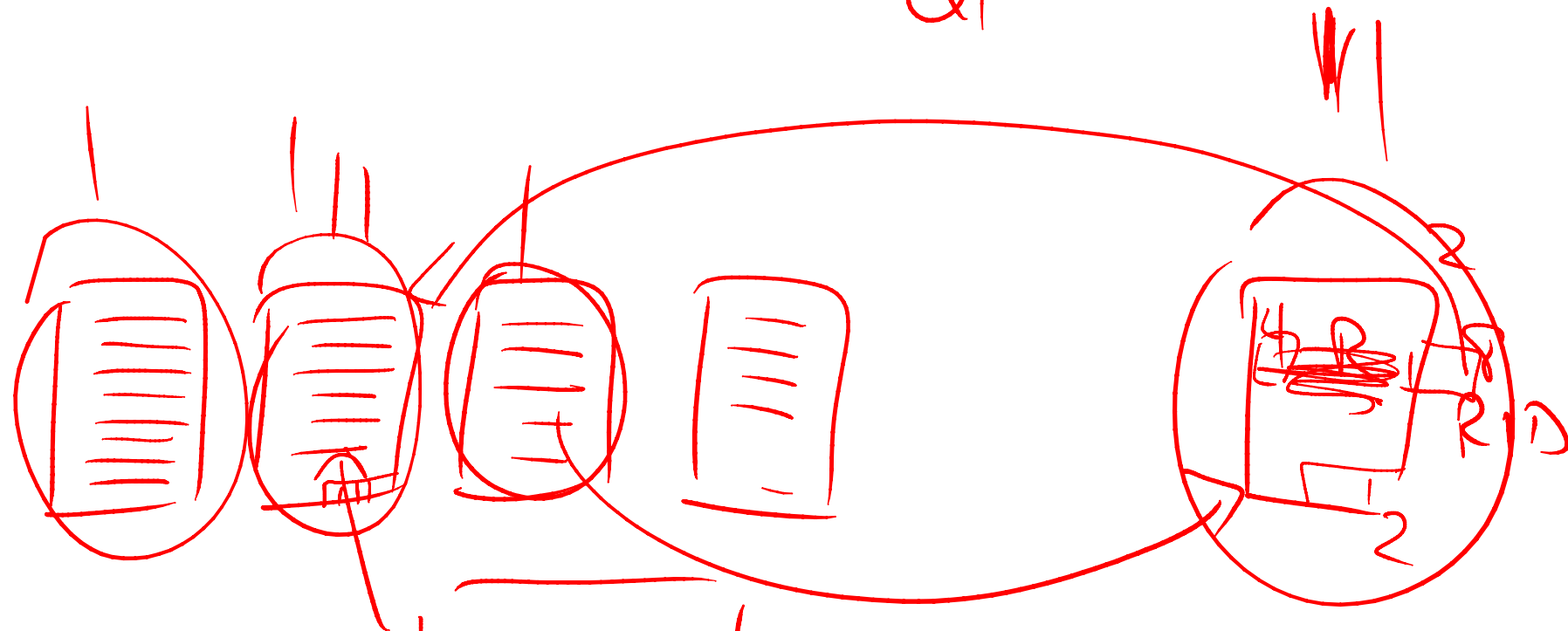
Each duplicate group can support $2^{31}-1$ duplicates. What happens if you put a CL index on Lastname and then have 2 billion rows with a value of Smith? Error 666.

Error 666: The maximum system-generated unique value for a duplicate group was exceeded for index with partition ID %l64d. Dropping and re-creating the index may resolve this; otherwise, use another clustering key.



Another negative associated with the uniquifier value is finding it. Remember, each duplicate has a uniquifier that's tied only to it. On insert, SQL Server will have to scan the last page to find the maximum *current* value for the uniquifier for that "group." If you have something like a DATE (not datetime/datetime2) then you're likely to have an impact from the generation of the uniquifier. It's often better to use something like an invoice number to unify the rows. And, secondarily this should also save time. Because the uniquifier has to live in the variable block it requires at least 6 bytes and possibly 8. These additional bytes are in the data rows and in ALL indexes. If 90% of your data rows include the uniquifier AND many of your nonclustered indexes include the invoice number then you're going to be require less storage with an int and probably even still with a bigint given that the table won't need the uniquifier and many indexes [probably] already have the invoice number.

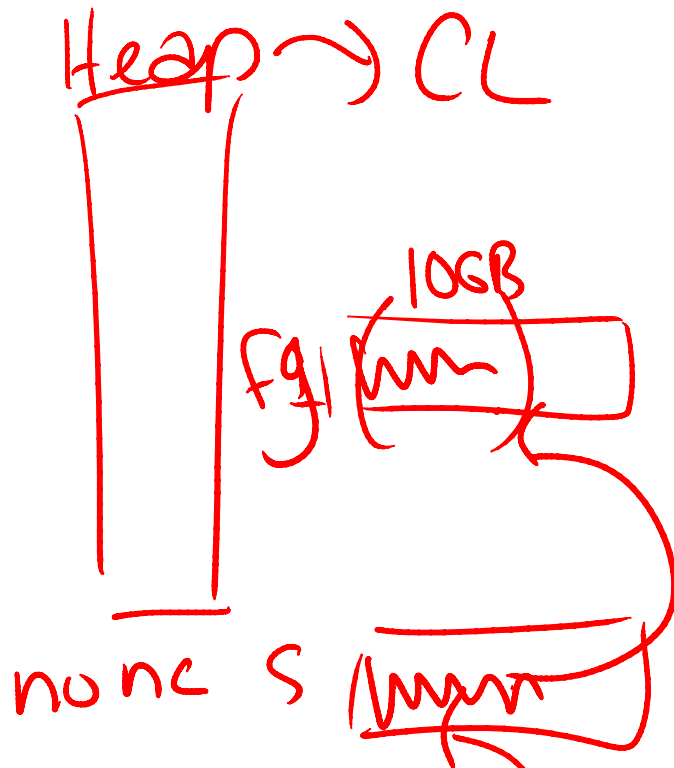
21



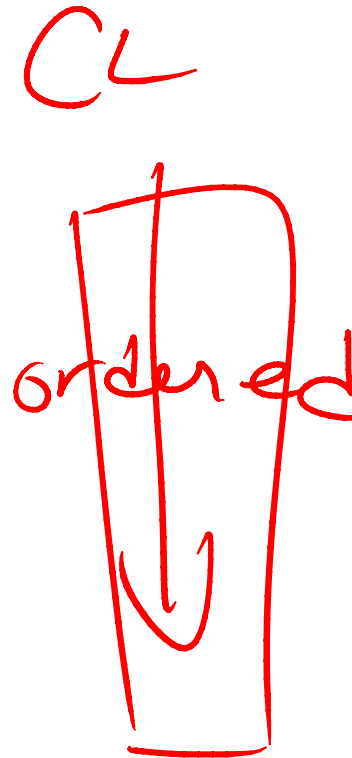
S12 – Forwarding pointers and back pointers (HEAPs ONLY)

This picture shows the overhead needed (in terms of bytes) for forwarding records. On the page where the record was inserted originally (to which ALL indexes will always point – as it's our “fixed RID”) will still have the header, the RID and the slot in the slot array (11 bytes). On the page where the row has been forwarded – you'll have a back pointer. This back pointer is 8 bytes but it lives in the variable block in the row so it requires an additional 2 bytes. Total space wasted is 21 bytes.

LOAD



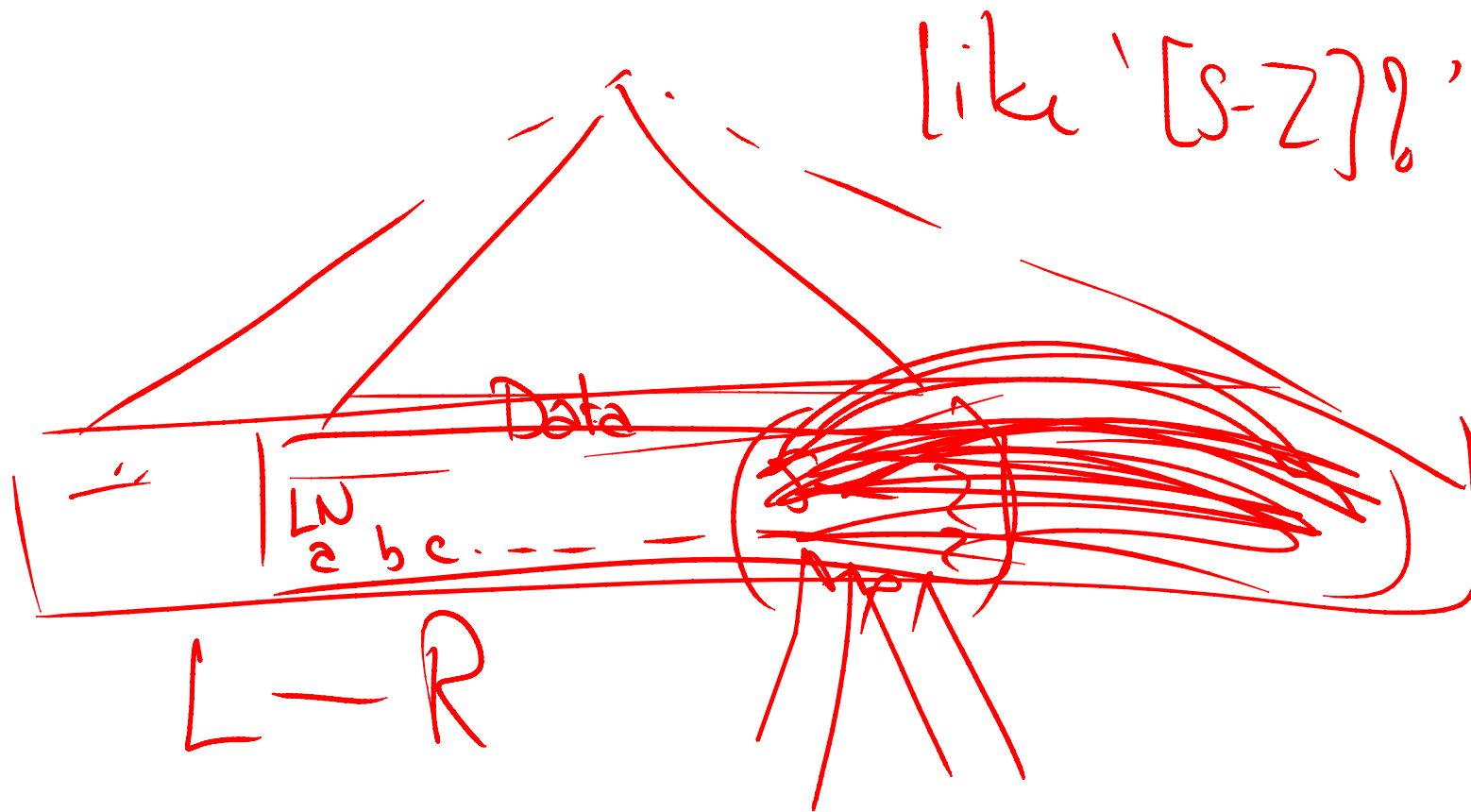
Load



High Performance Loading Discussion

Some really good questions about "high performance loading" that lead to a discussion around these key things:

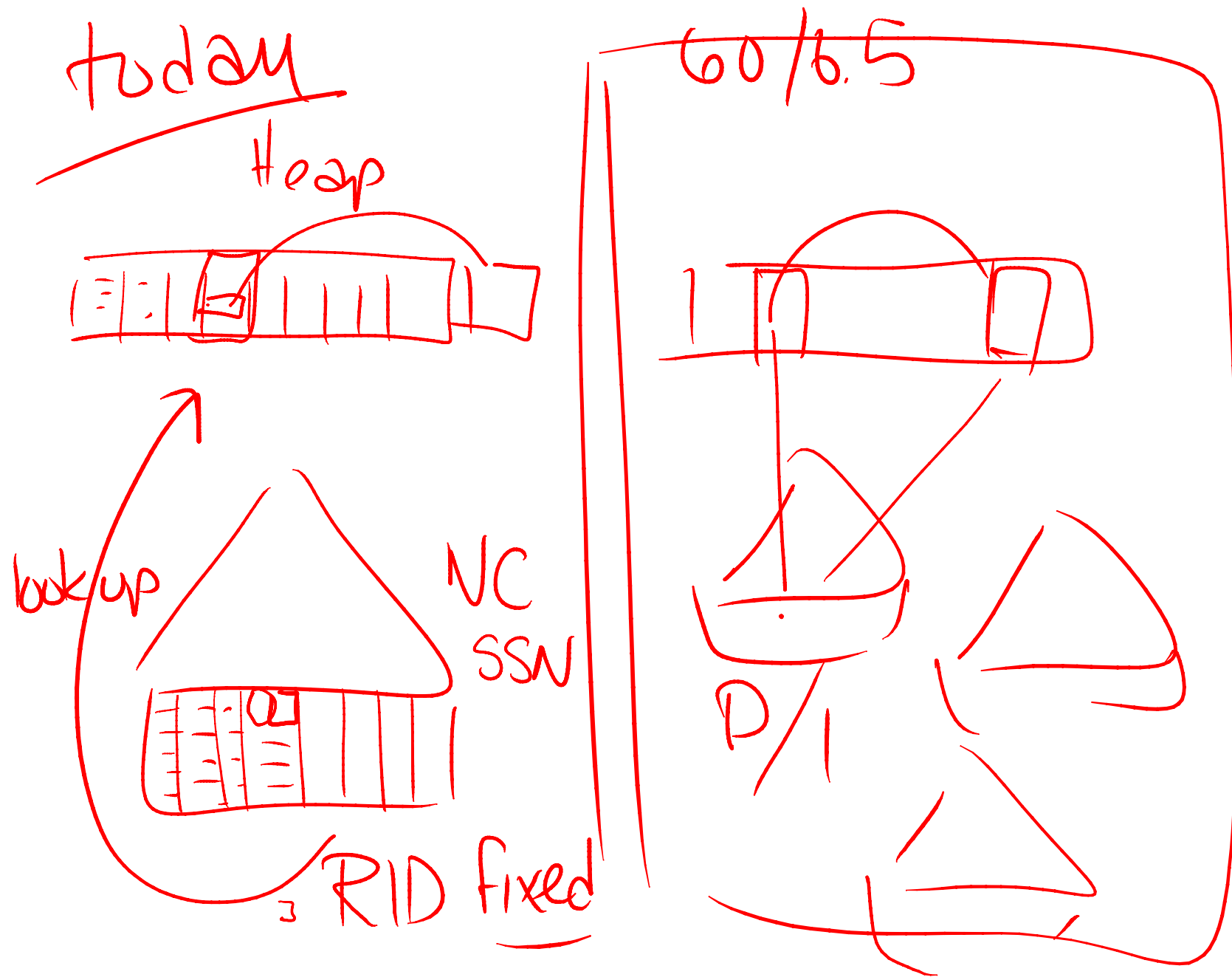
- **Loading into a HEAP** – typically this is best when you don't have an ordered file *and* when you can performance parallel load and parallel index creation.
- **Loading into a Clustered Table with an Ordered Load** – typically this is best when you have an ordered file and you cannot perform a parallel load and/or parallel index creation.
- **Loading into a Clustered Table with all nonclustered indexes pre-created** – this is almost NEVER a good idea unless you're loading only a small amount of data into an existing table. Typically less than 10% new data.



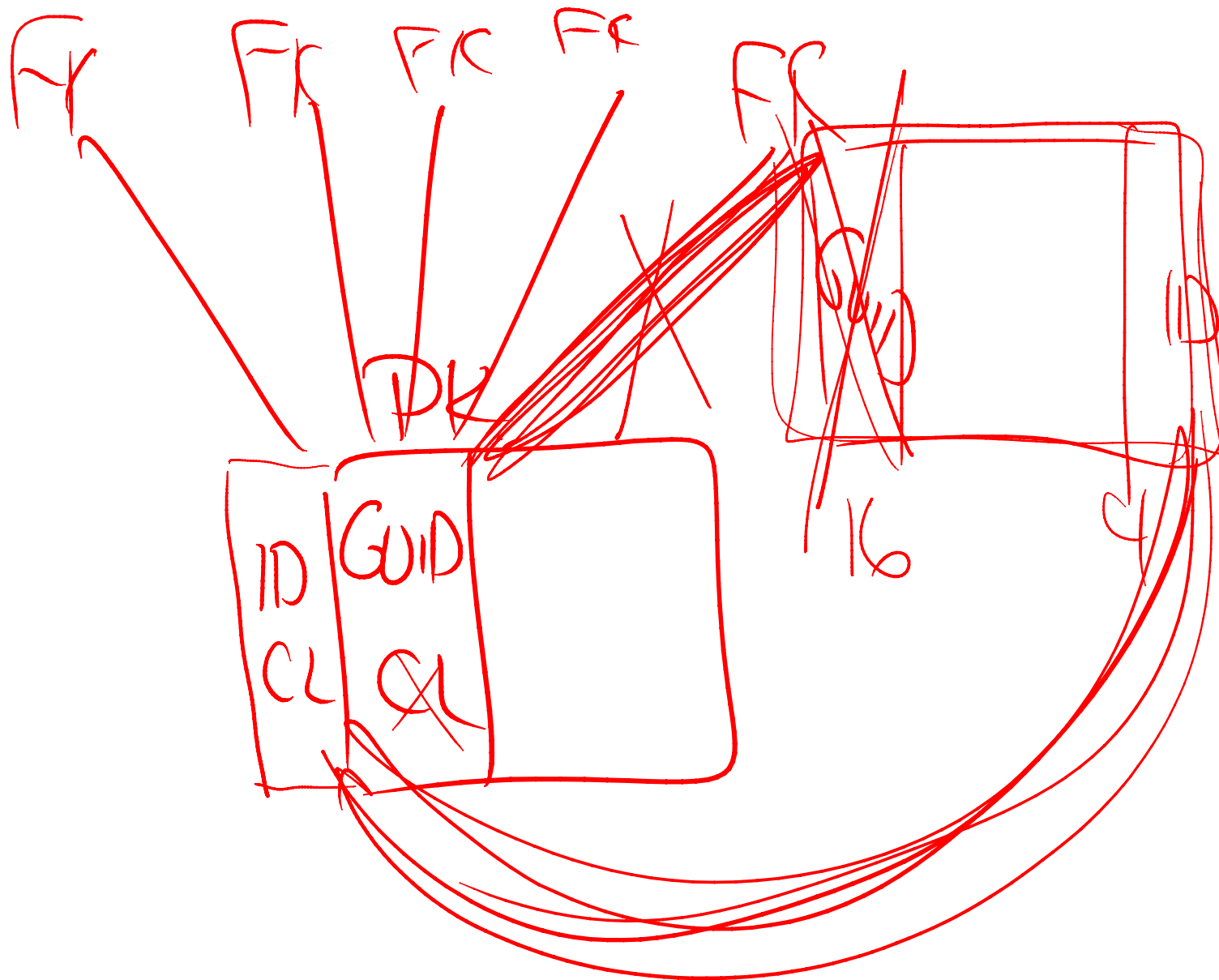
Clustering for range queries looks GREAT on paper... until modifications occur!

This was just reminding you that when folks tell you that the clustered index should be DESIGNED to support range queries you need to remind them of what the table is going to look like when there are modifications. What sounded good (the data is all together) doesn't stay that way...

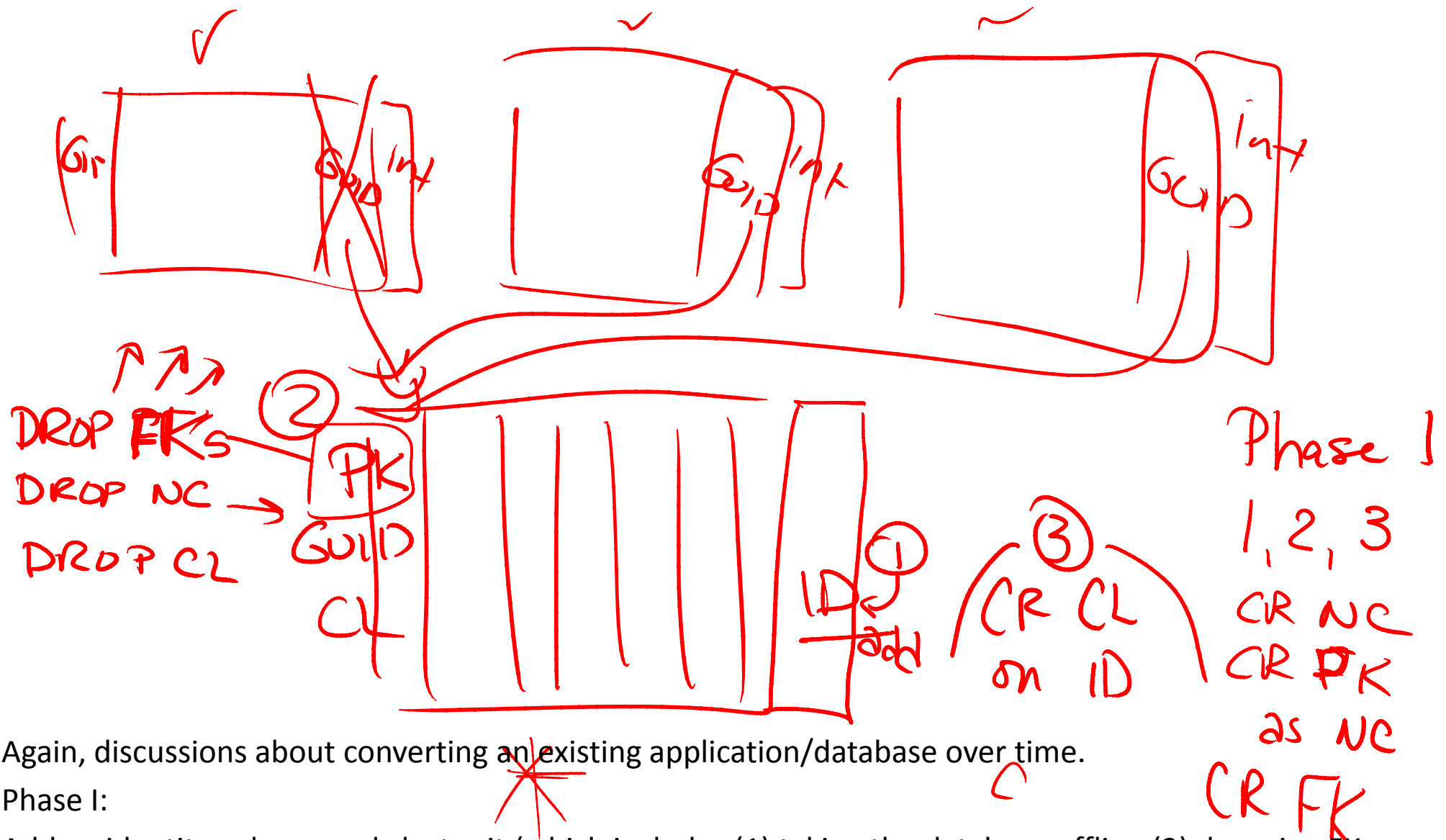
- INSERTs/UPDATEs are slower due to the splits
- This wastes disk space AND memory because of the fragmentation caused.
- And the range query is no longer left-right ordered but instead very out of order (fragmentation)



SQL Server 7.0's storage engine was almost a complete rewrite. A major change in the engine around the method used from nonclustered indexes to lookup the corresponding row in the table (heap or clustered). In 7.0 and higher, SQL Server uses a FIXED RID (for heaps) and the clustering key (for tables with a clustered index). In 6.x, SQL Server used a volatile RID for ALL tables (regardless of whether they were clustered or not).



Even if your “primary” table uses a GUID for inserts... You might want to use an IDENTITY column as your foreign key. This takes less data space, less index space and results in a less expensive join. And, this might be an easier “conversion” to make for your application. You can potentially make it over time and through changes to some tables, some procs and then some code. Then, more tables, more procs and other code, etc. Slowly you can remove the old GUID columns and eventually become free of GUIDs as clustering keys and GUIDs for joins! You might still have a GUID on that initial table but only as a nonclustered PK. 😊



Again, discussions about converting an existing application/database over time.

Phase I:

Add an identity column and cluster it (which includes (1) taking the database offline (2) dropping FKs, dropping NCs and finally, dropping the CL. (3) Add the identity column and cluster it. (4) Create the PK as NC. (5) Add back the other NCs. (6) Add back the FKs.

Phase II (you can do this for a subset of tables at a time, repeat until all are gone):

Migrate the related tables to use the Identity column for joins instead of the GUID-based FK. Slowly change code and eventually drop the GUID-based FK leaving ONLY the int-based FKs.



If you have only one file in a read-write filegroup you can potentially end up with contention on the system resources (PFS, GAM and SGAM). By having multiple files within the filegroup you can better handle contention.

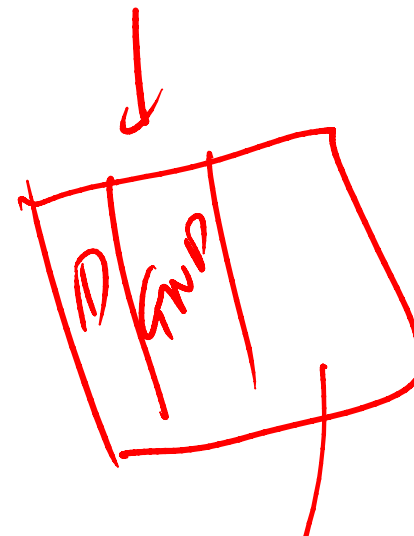
Generally, I recommend 2-4 files in a read-write filegroup. You do NOT need 1 per core.

File / pfs / gam / sgam
Global Shared Global

NC Ind Gender



Non-unique nonclustered indexes MUST have the lookup key (the Heap's RID or the clustered table's clustering key) pushed up the tree. The reason stems from the fact that the rows would NOT be able to be found on a delete. And, it guarantees their position on an insert.



Fixing GUIDs

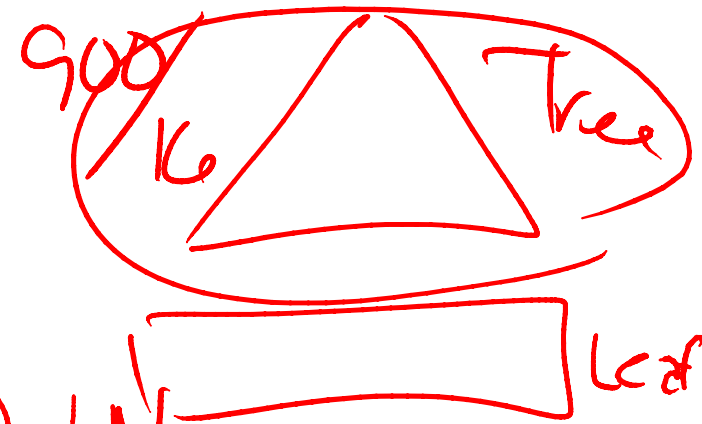
This scenario was from a question regarding the conversion of a database that uses GUIDs everywhere.

Completing eliminating GUIDs is tough to do. What you might do is something more gradual. Allow the application/user to continue to use the GUID as a nonclustered PK and then add a clustered identity. Slowly convert the related tables over by adding the identity column (as a FK back to the primary table) and then slowly remove this column altogether.

Table CL on EmpID

Key SSN
Inc LN

Key SSN
Leaf SSN, EmpID, LN



non-unique index on phone

Key phone
Inc LN

Key phone, EmpID
Leaf phone, EmpID, LN

S45

Where does the CL key go?

If the index is unique – then the CL is added only to the leaf-level.

If the nonclustered is non-unique – then the CL is added to leaf level as well as up the tree.

Specifically, it's added immediately after the key; included columns are added after that.

SQLskills Immersion Event

IE1: Internals and Performance

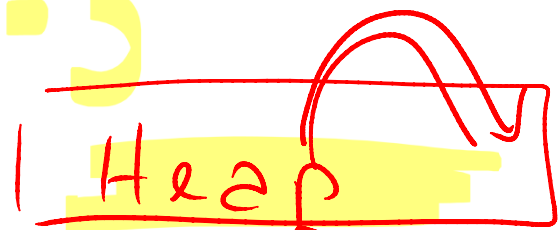
Module 9: Internals and Data Access

Kimberly L. Tripp
Kimberly@SQLskills.com



4-5 10

2-3



1.6 million

~10K

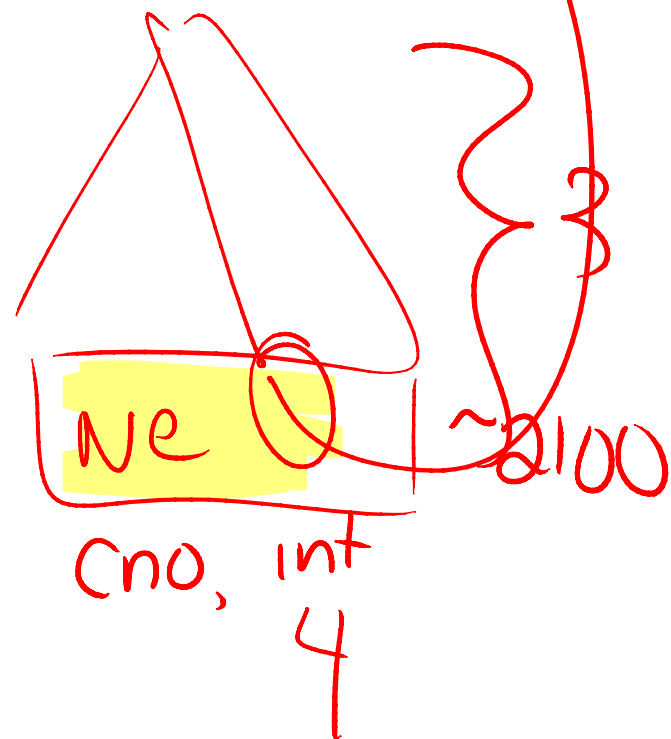


2

~pk



NC
Charge-no



M8 Question

Discussing the costing (of IOs) for bookmark lookups – between a heap and clustered table.

It *looks* like the CL index is worse:

Using a nonclustered to do a lookup (= 3 IOs), then – using the CL to lookup the data row = 3 more IOs for a total of 6 IOs

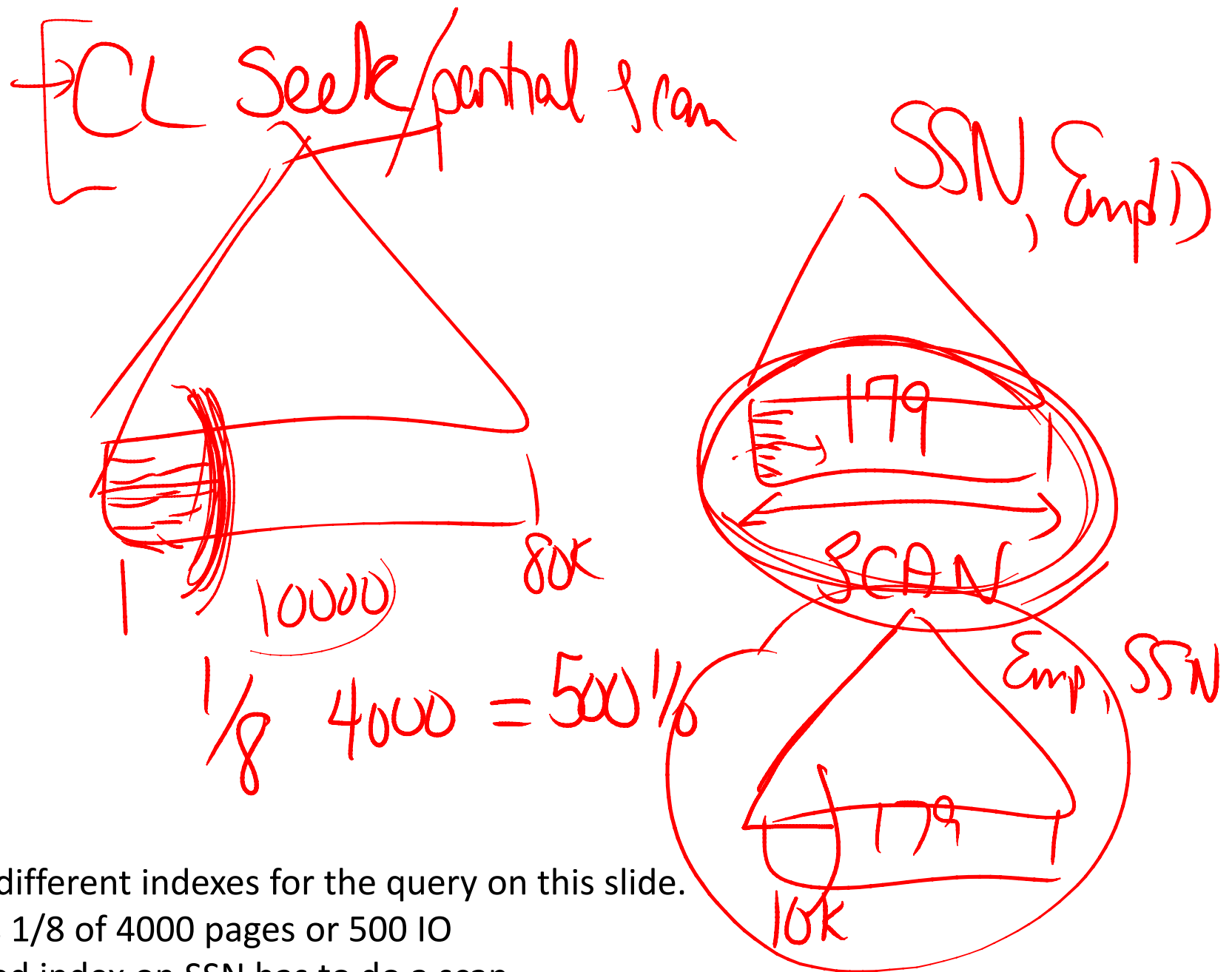
Whereas the heap seems to require fewer IOs. Using the nonclustered is about the same (=3 IOs), then – using the Heap to access the data row is 1 (possibly 2 if there's been record relocation) for a total of 4-5 IOs.

The long story short is that 4-5 IOs seems less than 6 IOs. Yes, the number is less but the IOs are potentially more expensive. The yellow highlighting show where the more expensive IOs are going to be performed (which is predominantly in the leaf structures).

As a result, a bookmark lookup from a NC to a clustered has 2 potentially physical IOs.

Whereas the lookup from a NC to a heap has 2-3 potentially physical IOs.

While many lookups might be the same – there are still OTHER reasons for why heaps are not ideal. This is just yet-another-one. ☺



S16

Costing IOs for different indexes for the query on this slide.

The clustered is 1/8 of 4000 pages or 500 IO

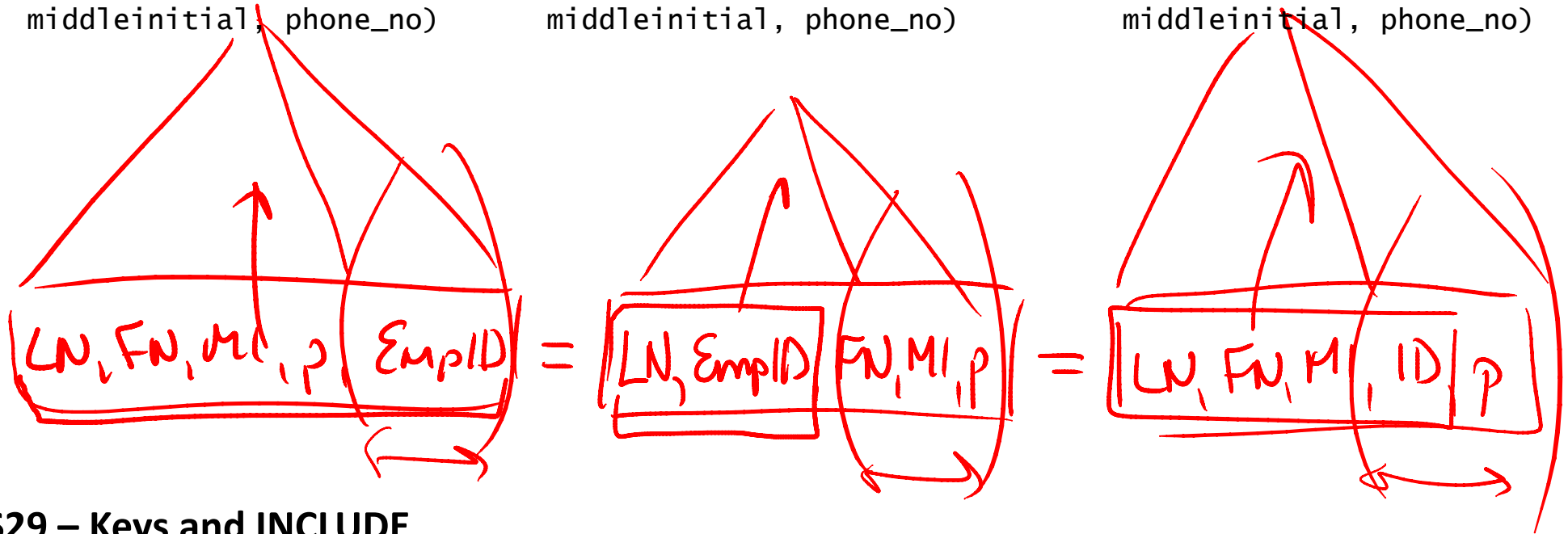
The nonclustered index on SSN has to do a scan

But, a nonclustered covering index that was ordered EmpID then SSN (S17) would be a seekable nonclustered covering index with the fewest IOs as SQL Server would only have to read 1/8 of the 179 pages.

```
CREATE INDEX NCICovers4Cols
ON dbo.member
(lastname, firstname,
middleinitial, phone_no)
```

```
CREATE INDEX NCILNinKeyInc3Cols
ON dbo.member(lastname)
INCLUDE (firstname,
middleinitial, phone_no)
```

```
CREATE INDEX NCICovers4Cols
ON dbo.member
(lastname, firstname,
middleinitial, phone_no)
```

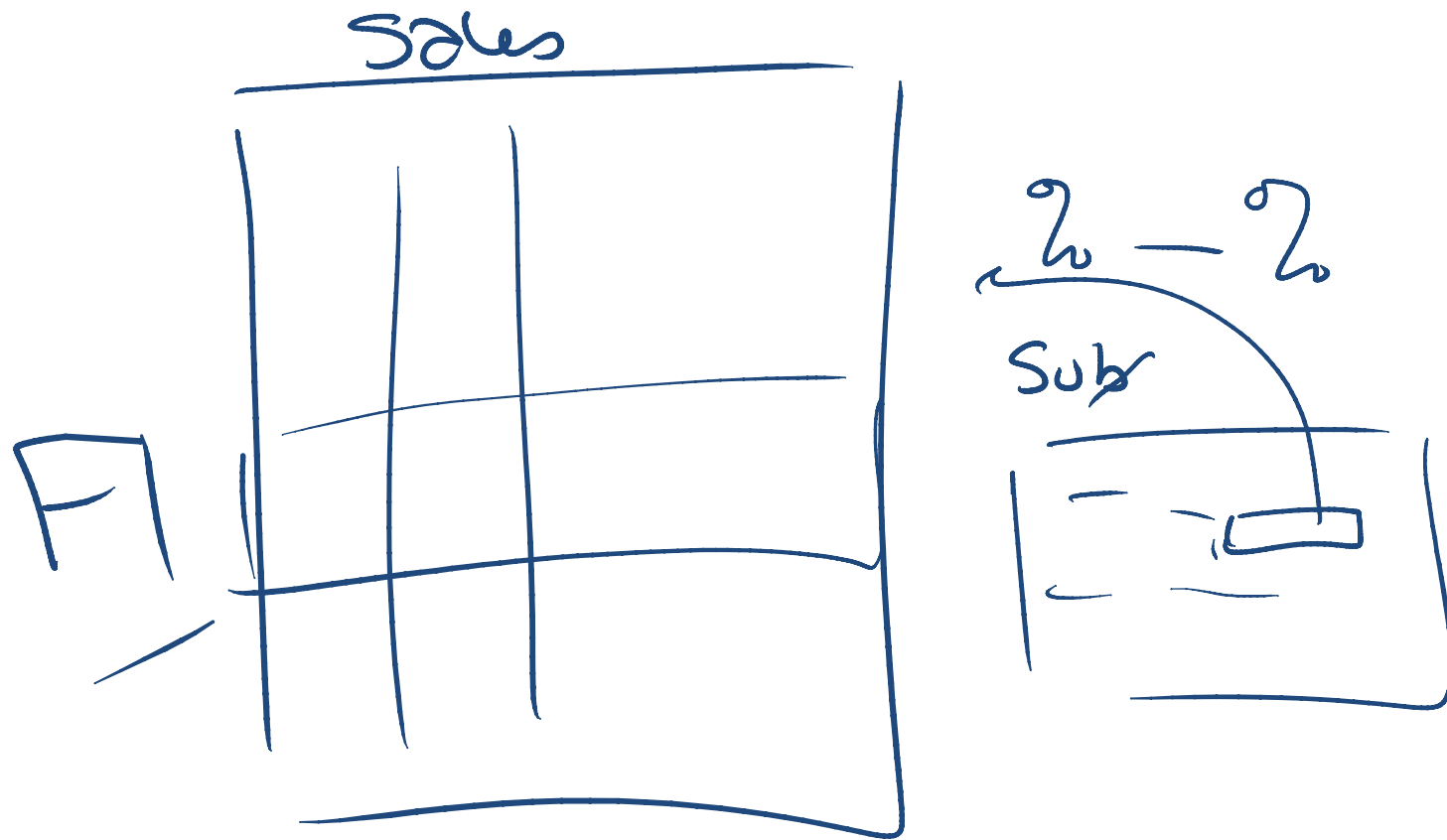


S29 – Keys and INCLUDE

In the first example all of the columns are in the key. The clustering key is added to the leaf-level and then all 5 columns go up the tree.

In the second example only lastname is in the key. The clustering key is added immediately after that (in the leaf-level) and then the included columns are added to the leaf-level. Then, only lastname and EmpID go up the tree. This is the best index for this query (if we were to tune query-by-query, which is completely unrealistic).

In the final example phone is the only include. The clustering key is added immediately after LN, FN, MI (in the leaf-level) and then phone is added. LN, FN, MI, EmpID go up the tree. This last example is the most realistic index but only after consolidation warrants it.

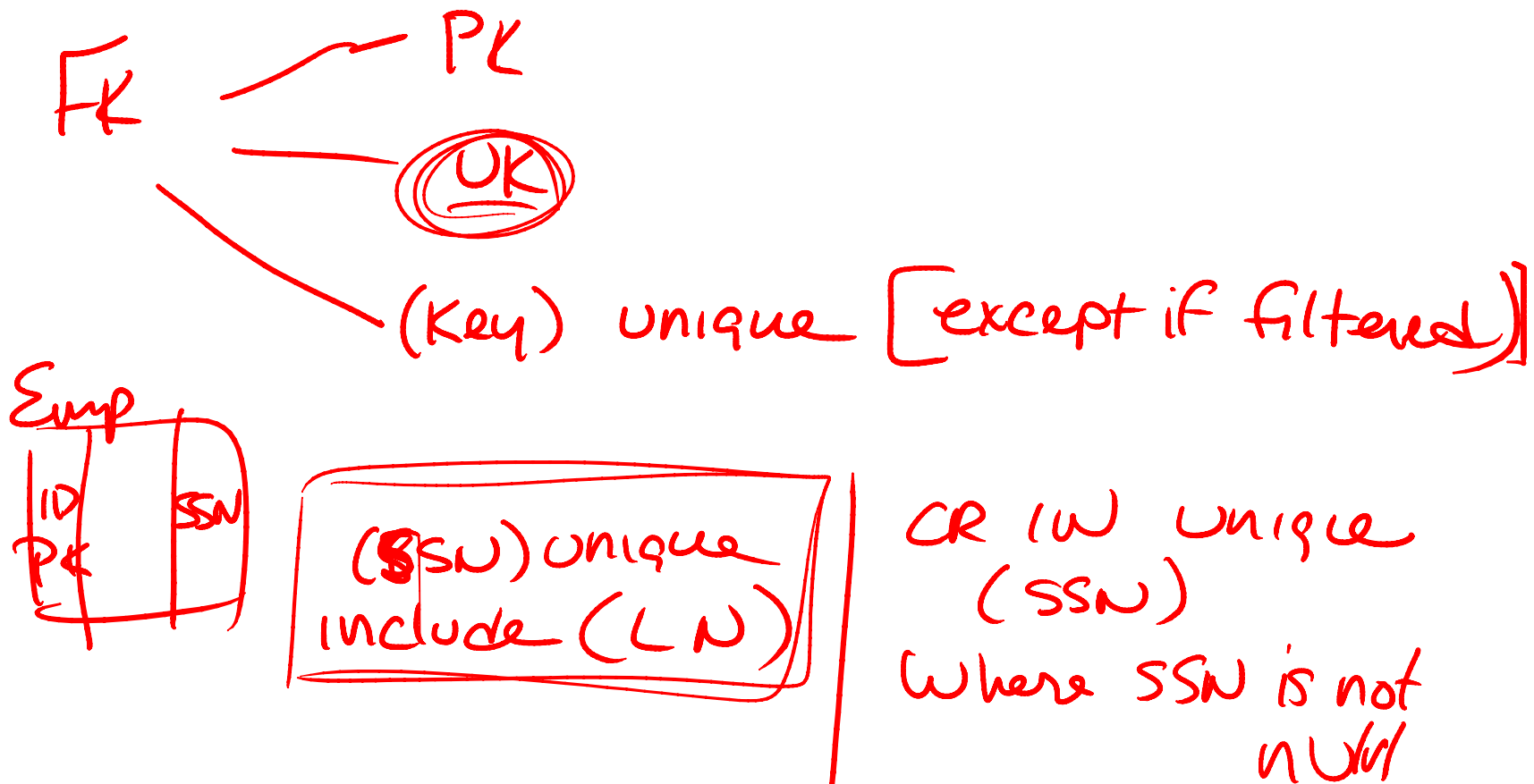


Solving an application problem with a filtered index

Somewhat recently I worked with an application that had a habit of adding % to the start/end of a queried value but only for a subset of data. However, that subset was not selective enough to use a nonclustered index. So, the query would do a complete table scan. This was causing huge performance problems... enter, filtered index!

We could cover the request but only where REGION = 'subset' so only that data would be scanned. OK, no, I didn't like the %value% search either but at least I could reduce the impact of it by using filtered indexes!

The negative – consistent session settings are required AND procedures must do a recompile to ensure that they will pick up the filtered index.



A foreign key can be used to reference any column (or list of columns) that has a unique index – as long as it's not filtered.

In an index consolidation scenario, if you already have a constraint on SSN and you NEED an index on SSN include (LastName) you'd end up with a "redundant" index (if SQL *required* a constraint-based index for foreign keys). Instead, you can drop the constraint on SSN and instead create:

```
CREATE UNIQUE NONCLUSTERED INDEX SSN_Inc_LN ON Member (SSN)
INCLUDE (LastName)
```

Foreign Keys can reference UNIQUE indexes (without constraints)

[http://www.sqlskills.com/BLOGS/KIMBERLY/post/Foreign-Keys-can-reference-UNIQUE-indexes-\(without-constraints\).aspx](http://www.sqlskills.com/BLOGS/KIMBERLY/post/Foreign-Keys-can-reference-UNIQUE-indexes-(without-constraints).aspx)

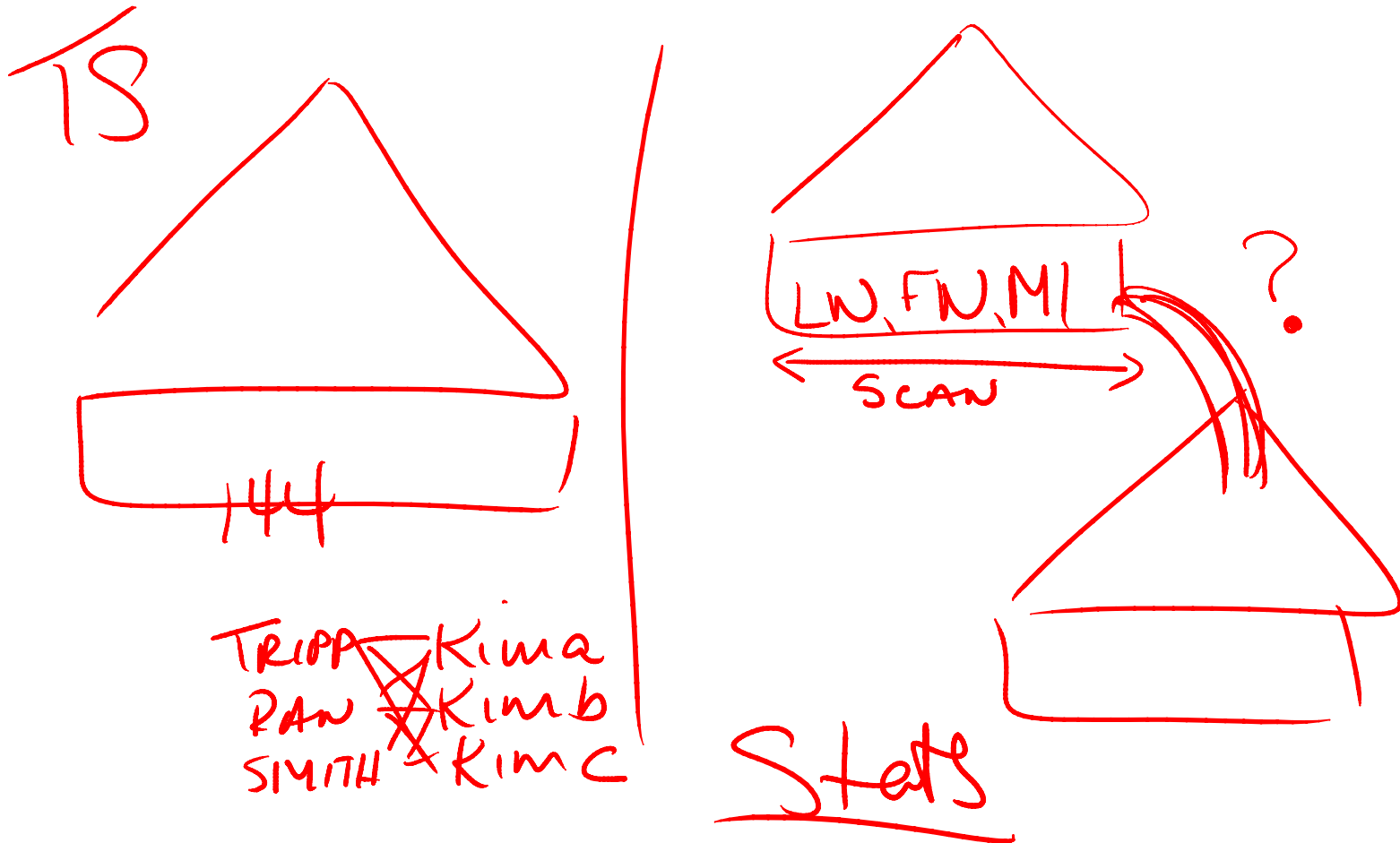
SQLskills Immersion Event

IE1: Internals and Performance

Module 10: Statistics

Kimberly L. Tripp
Kimberly@SQLskills.com





S11 – Column-level distribution from left-based density subsets...

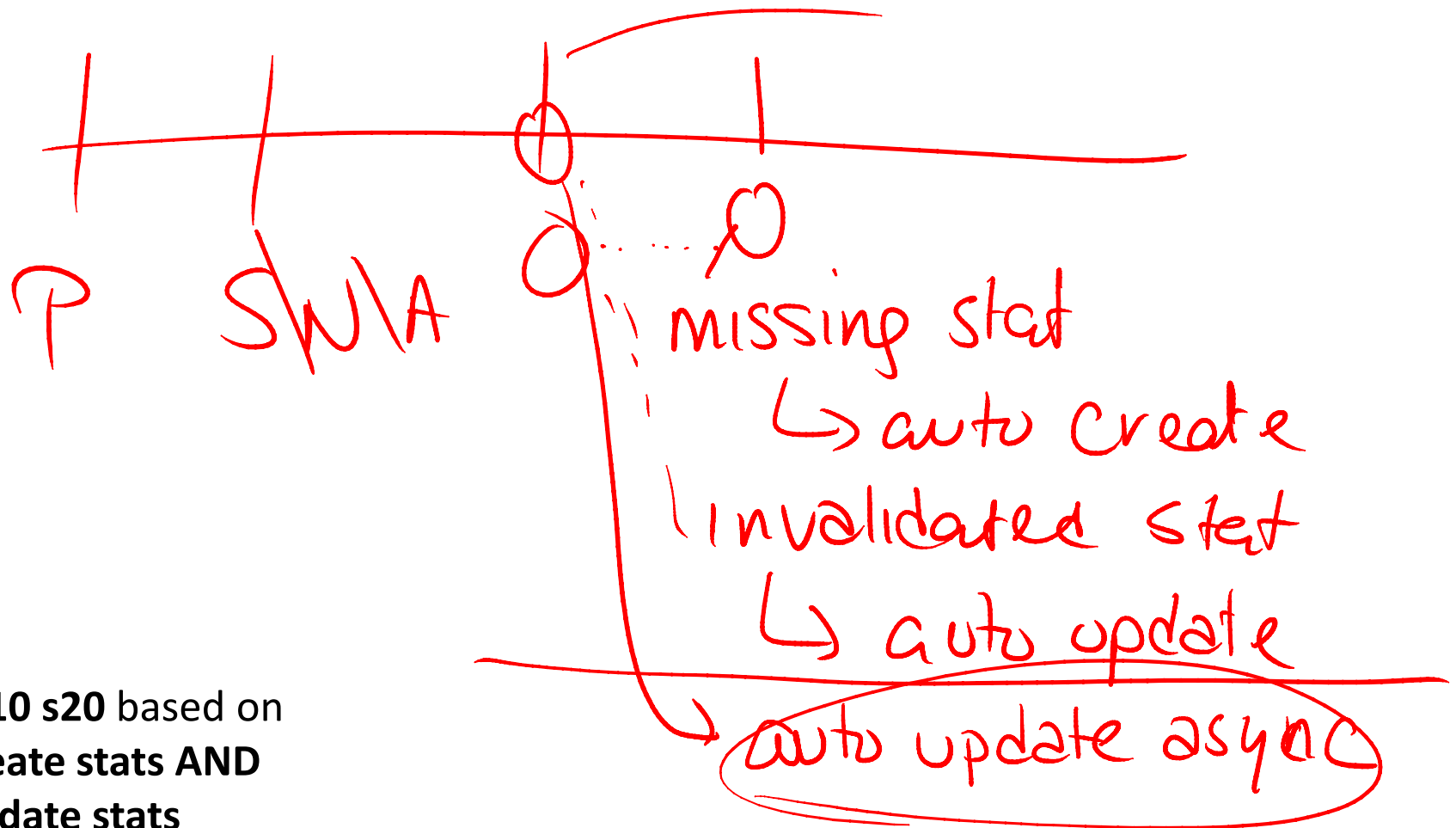
This relates to the query on slide 10 and the idea that I was trying to show here is that even though the left-based density shows that last names are horribly NOT unique and the combination of last name & first name IS *almost* unique – that doesn't imply anything about first names alone. I could have created a data set of firstnames of Kima, Kimb, Kimc and then multiplied that with last names (Tripp, Randal, Smith) and I would have had similar statistics for last name alone and for the combination of last name, first name. SQL Server does NOT gamble on this – SQL Server creates column level statistics on first name.



S15 – autocreate stats (same as prior slide)

This is just the reminder of what happened during the demo. When the query ran with `FIRSTNAME LIKE 'Kim%'` and SQL Server did not have enough information AND auto create stats is ON – then, SQL Server creates the statistic, optimizes and executes.

The process of optimization actually stops while the statistic gets created. There's another summary of this on slide 20.



**New slide M10 s20 based on
S15 – autocreate stats AND
S18 – autoupdate stats**

During optimization – SQL Server might hit one of two events:

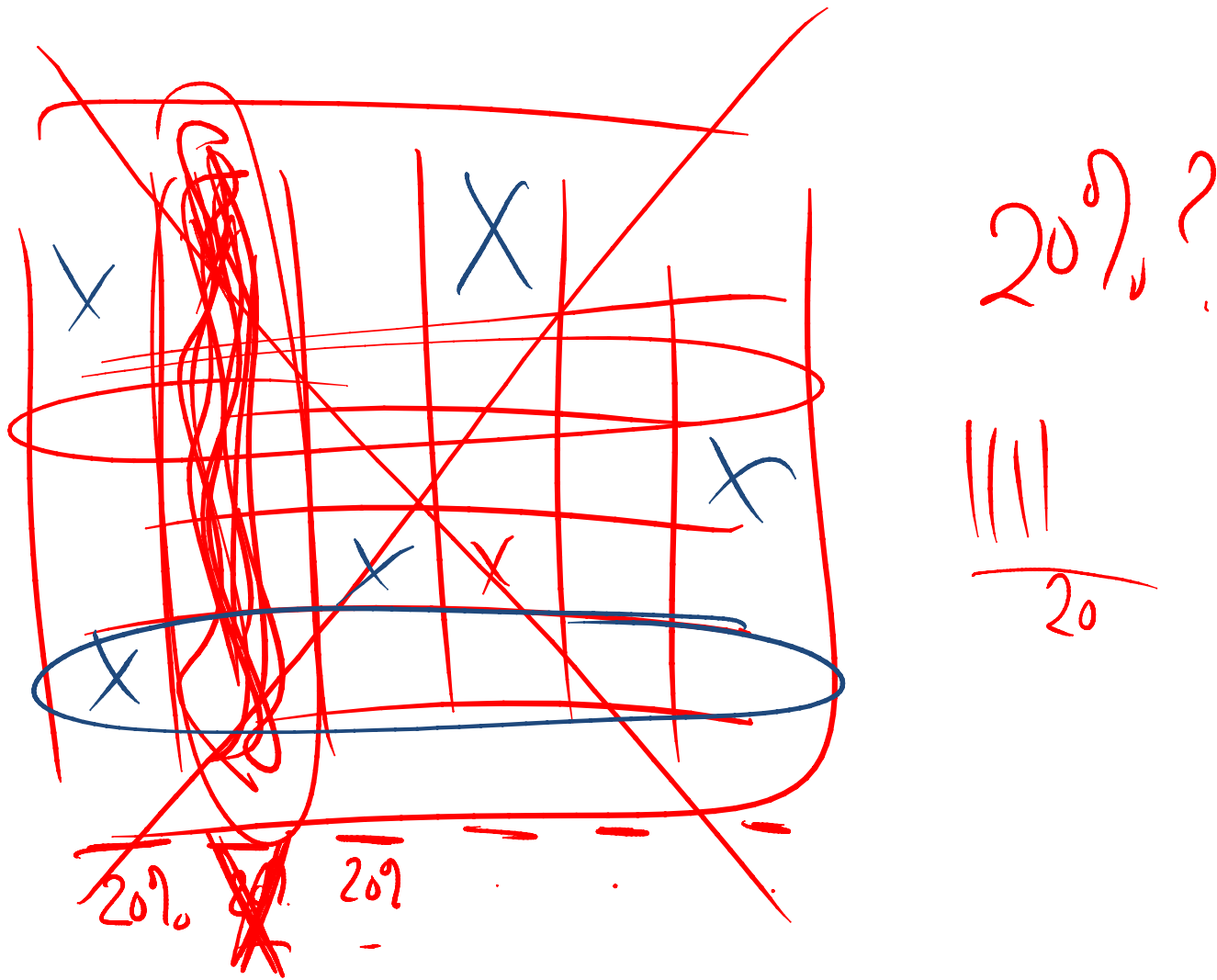
Missing statistics event

If “auto create stats” is on then the optimizer will wait until the statistics have been created to optimize.

Invalidated statistics event

If “auto update stats” is on then the optimizer will wait until the statistics have been updated to optimize.

If “auto updates stats async” are on then SQL Server will optimize using the invalidated statistics. Then, asynchronously, SQL Server will update the stats.



S19

This is a diagram showing the pros/cons of the methods for statistics invalidation. SQL Server 7.0 and 2000 used a rowmodctr (sysindexes.rowmodctr) to do invalidation. Even though SQL Server doesn't use this anymore – you can (through the backward compatibility view sysindexes). In 2005+ they moved to an internally defined colmodctr:

- The pro is that you don't invalidate too soon (when you have a highly volatile column)
- The con is that you might not invalidate soon enough if your modifications are reasonably distributed.

in

X				X		X	
		X			X		X

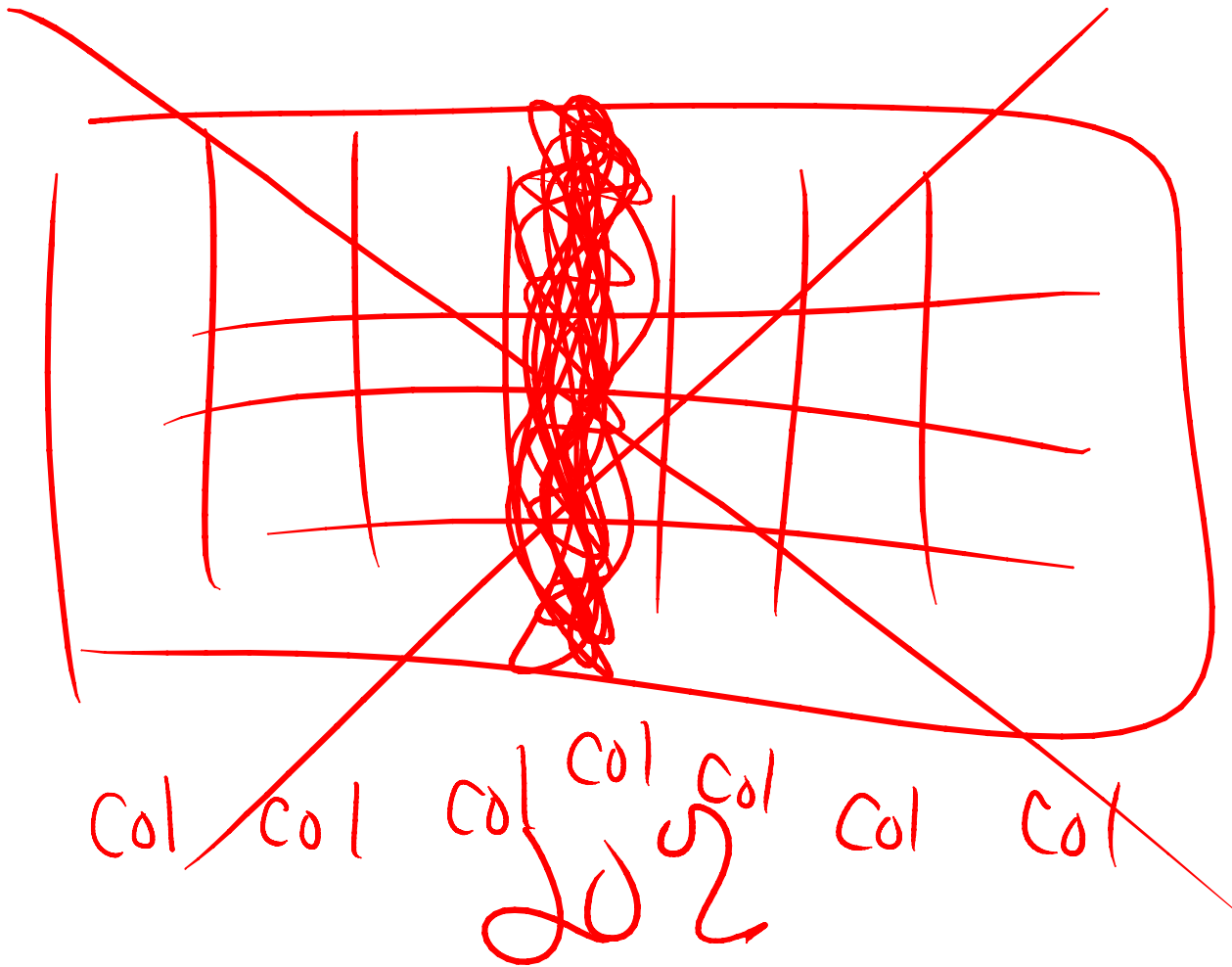
row modified 20%

|||||

S19

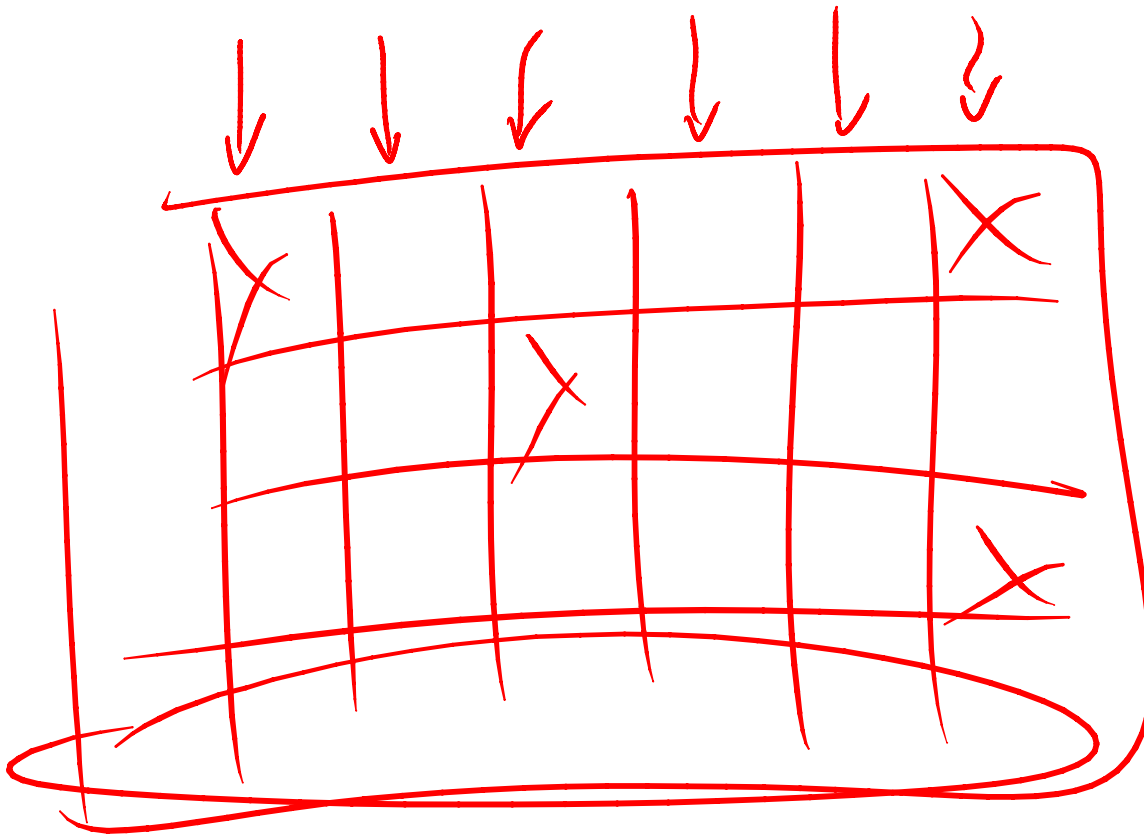
This shows how relatively distributed modifications effect each column with a small percentage of the modifications but when they add up to 20% ALL columns become invalidated (this was the PRE-2005 way of doing it):

- The pro was that each column had a reasonable (but lower) percentage of rows modified and the stats were invalidated
- The con is that a single overly volatile column would cause ALL statistics to be invalidated (which was overkill).



S19

The problem prior to 2005 was that a single overly volatile column would cause ALL statistics to be invalidated.

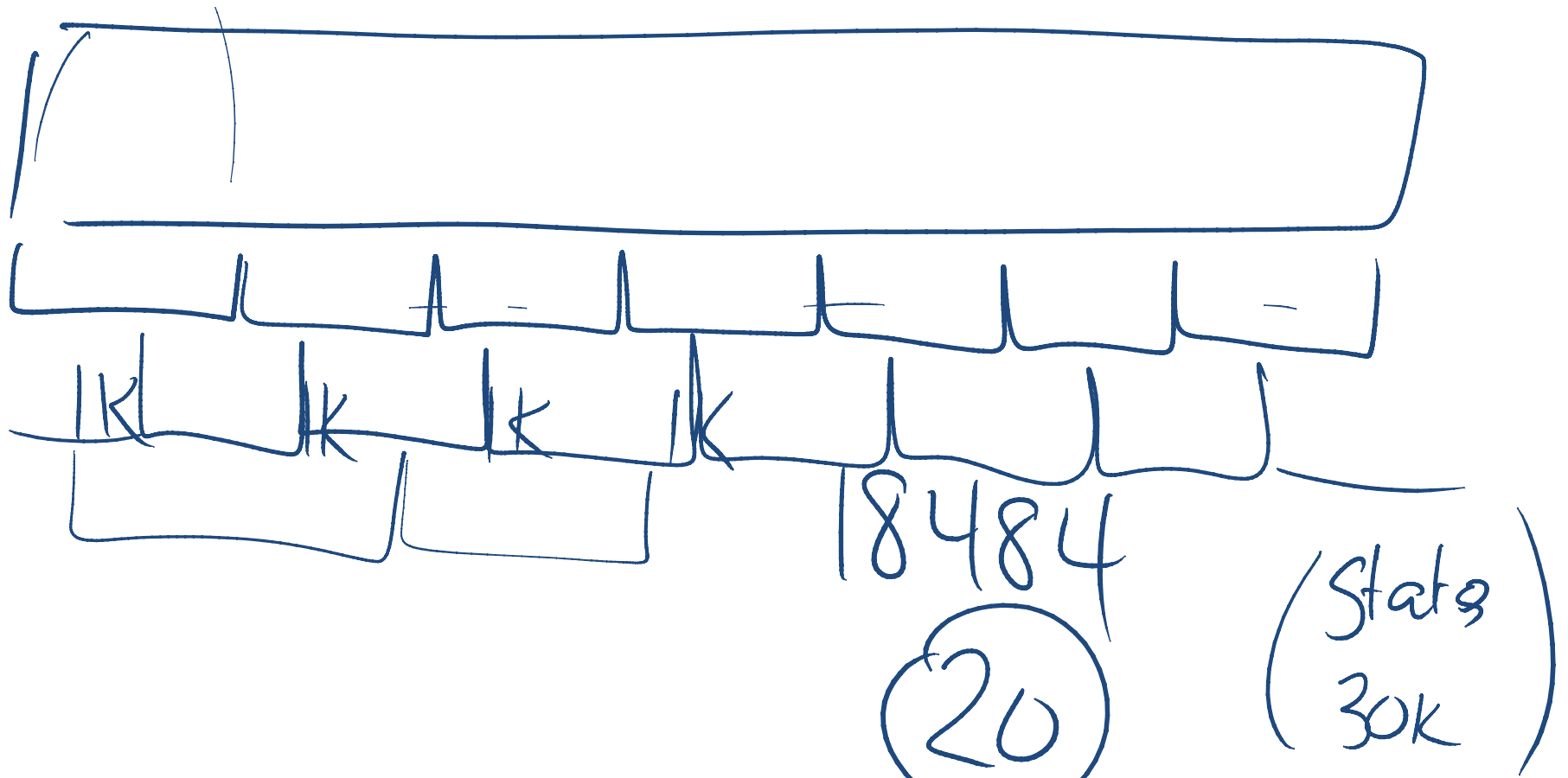


20%

S19

In 2005+ they use an internally defined colmodctr:

- The pro is that you don't invalidate too soon (when you have a highly volatile column)
- The con is that you might not invalidate soon enough if your modifications are reasonably distributed.

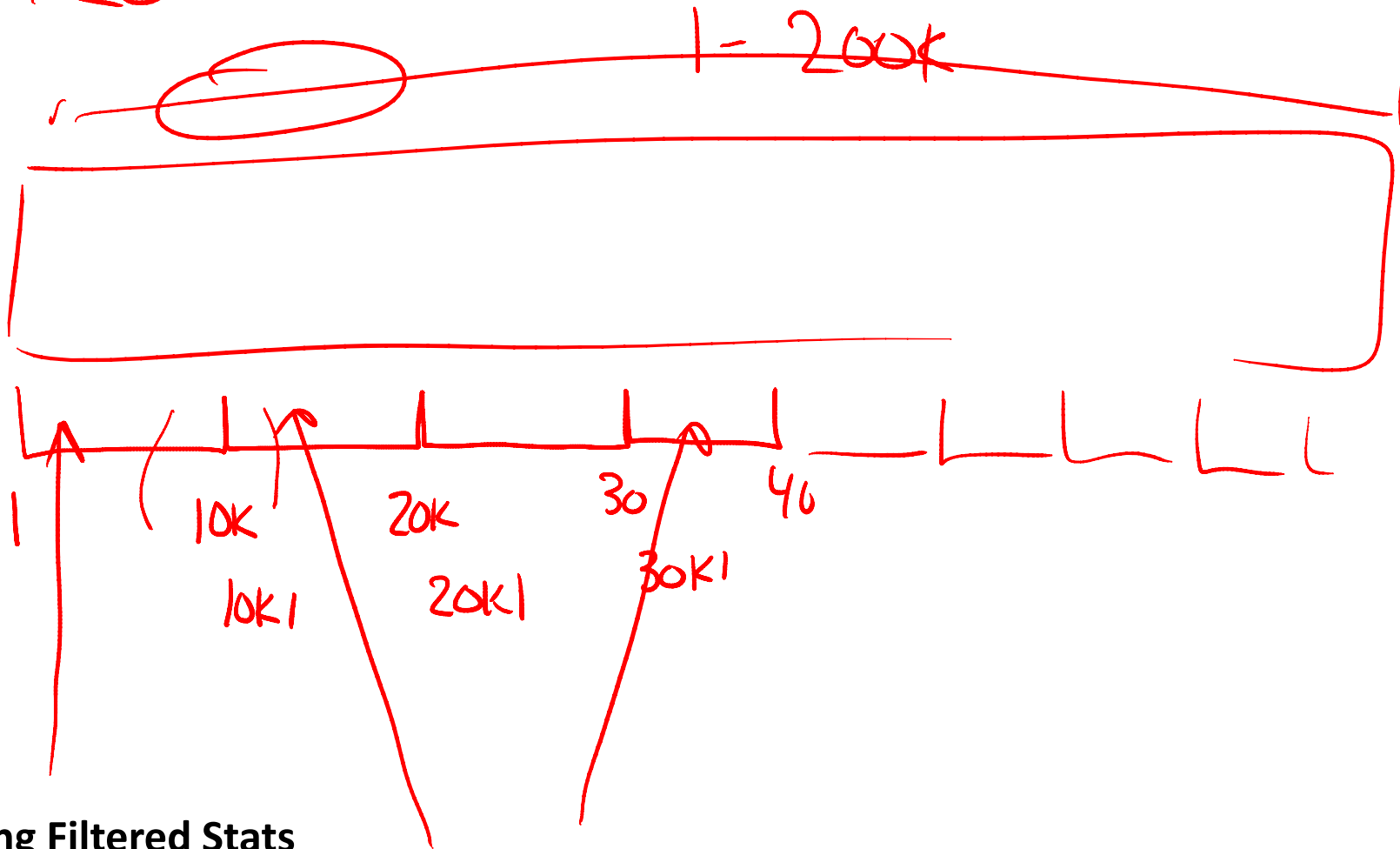


S28

Filtered statistics can be created for specific values but depending on your data distribution – you might want to divide the data into buckets and then create a [filtered] statistic for each of those buckets. The example was 31 million sales over ~18K customers. By creating a statistic for each 1K customers you have statistics that are effectively 19x more detailed. Even just 10 buckets gives you 10 times more detail. However, it still might not be detailed enough. You'll need to test it and check the histogram. Because of potential **interval subsumption*** issues– some have asked if it would be beneficial to create additional stats at different intervals... you could but it would really depend on the queries. Having said that – the bigger the range that the query is interested the more the averages are more accurate. So, really, this issue is to significantly help queries that are more targeted (where the stats just weren't good).

* The optimizer can detect whether interval conditions in a filtered index cover, or "subsume" interval conditions of a query.

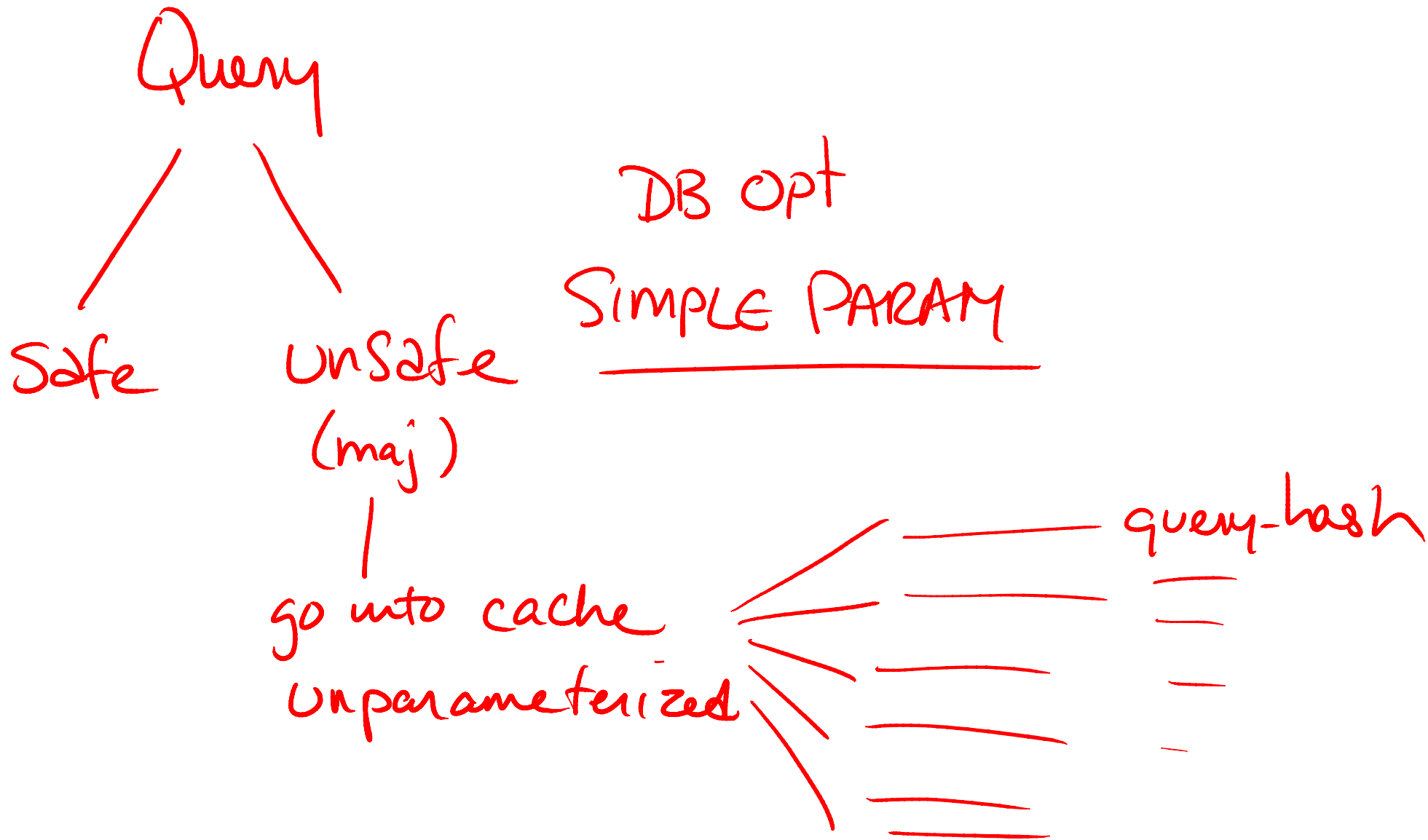
1/20 of data 200K cost 10K



Creating Filtered Stats

The idea is to just have better statistics than what you have currently. To do this you can divide the range into 10-20 buckets. This will give you 10 times (or 20 times) better information in terms of statistics.

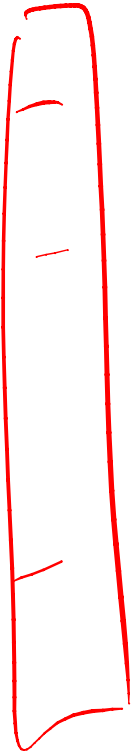
You WILL need to regularly/automatically review the values/ranges and add more or divide them up again to ensure that the statistics are good!



Simple vs. Forced Parameterization

The general process is that SQL Server analyzes a query to determine if it's safe or not (the majority of them will **NOT** be safe). If it's safe then it can get parameterized and reused. If it's unsafe then it goes into memory as an individual query (and it's harder to determine the cumulative effect of these queries). Check out the query_hash and query_plan_hash.

table



Status = 1

filter = 1

Query



Where Status = 1

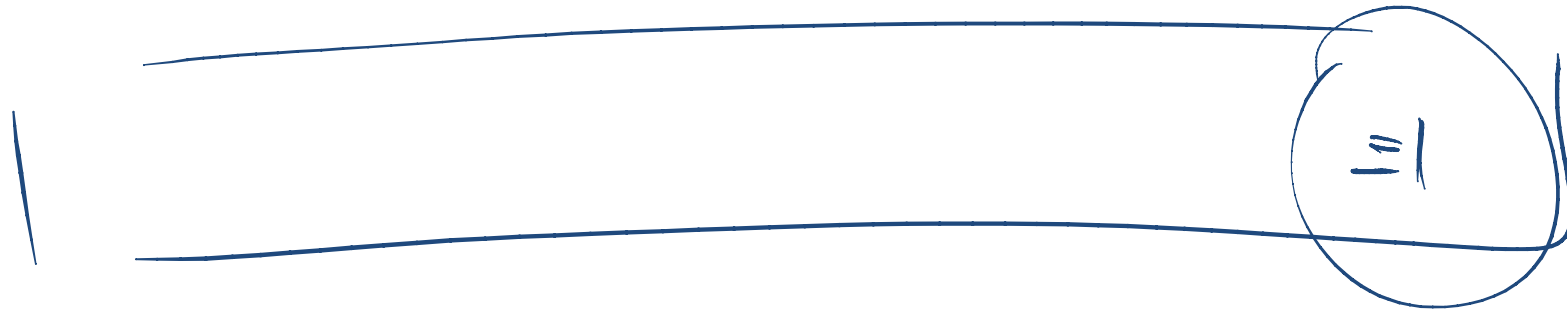
Simple analyze use index

Forced status = @1

S Filtered Stats and Forced Param

The bad news... there's always bad news.

Forced parameterization: even when SQL Server would NOT have parameterized the statement (because it had deemed the statement's plan as "unsafe" to re-use), forced param will force it. In systems where plans are very stable (but A LOT of adhoc) then this could be great. However, the example of status = 1 (in your query) is converted to status = @1 so the filtered index/filtered stat cannot be used.



Simple

Status = 1 not
save plan

FORCED

Status = @1

S28

The bad news... there's always bad news.

For Updates:

statistics for a filtered index OR a filtered stat – do NOT get updated until that column's statistics get updated (which is when the colmodctr) is reached

Forced parameterization: even when SQL Server would NOT have parameterized the statement (because it had deemed the statement's plan as "unsafe" to re-use), forced param will force it. In systems where plans are very stable (but A LOT of adhoc) then this could be great. However, the example of status = 1 (in your query) is converted to status = @1 so the filtered index/filtered stat cannot be used.

F1
F1
F1
F1

~~Status = 1~~

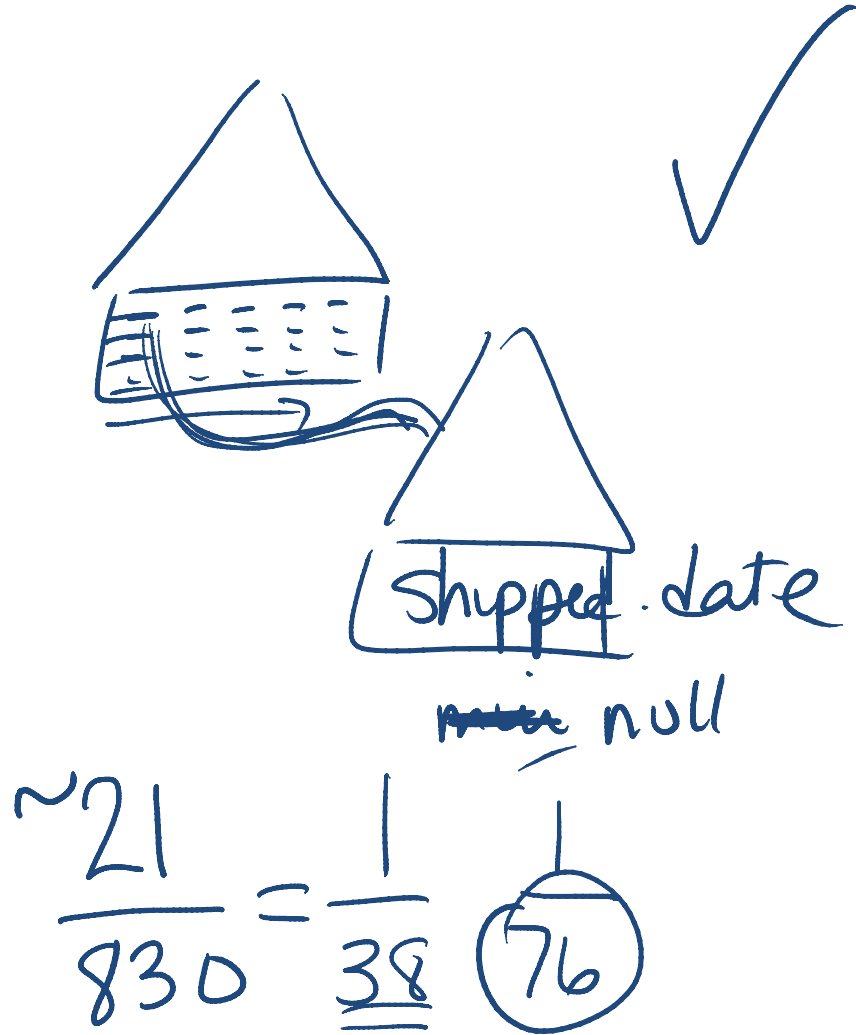
Status = @1

Database option: FORCED parameterization

Because of how forced parameterization changes each variable to a parameter – the value of that parameter is unknown. As a result, a filtered index (for example, WHERE status = 1) cannot be used when a query has been parameterized to status = @1. There's no guarantee that the value they're searching on is 1.

End result. You will NOT have good results with filtered indexes IF you use FORCED parameterization. **The good news:** This is NOT on by default and not likely to be on in most environments.

Index on OrderDate



Index on ShippedDate

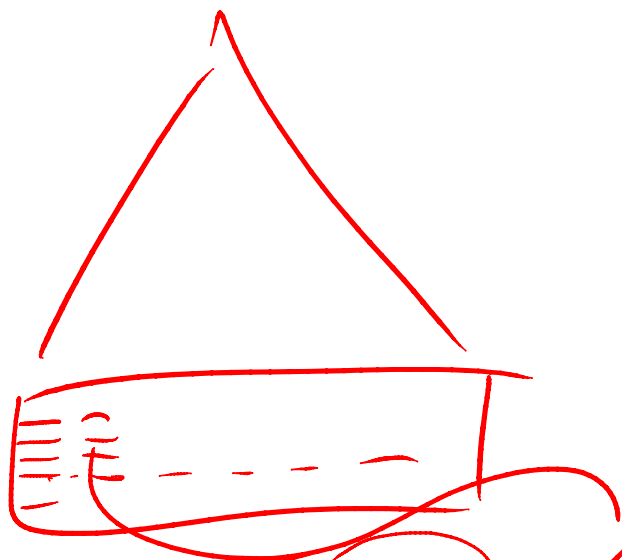


S30

This is the diagram to show the estimated costs using an index on OrderDate and an index on ShippedDate. The text on the slide describes the requirements.

TS
OD

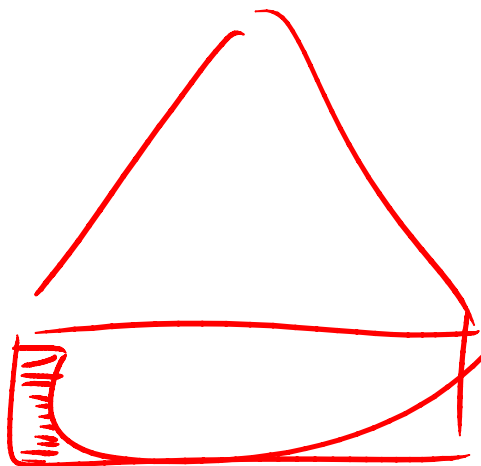
$$21/830 = \cancel{38} 76$$



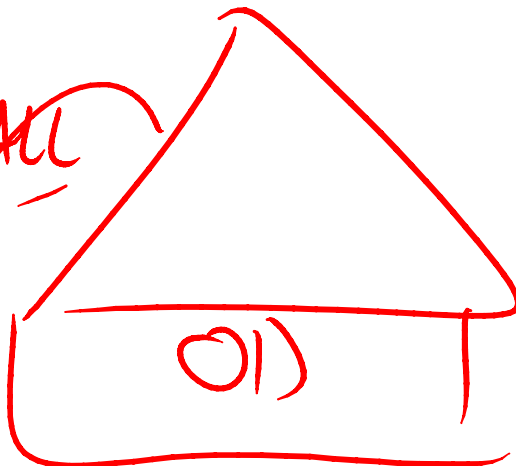
lookup
 ≤ 76



SD



ALL



sort

Poor Man's DW

OLTP → B/R

① Update stats ←

② Rebuild Indexes

FF = 100

~~Update stats~~ ←

Sampling?

③ sp_createstats

Index only

Full scan - ?

READ ONLY

Poor Man's Datawarehousing

Setting a database to RO (read-only) can be a good idea when you plan to use it solely for reads. However, how does SQL Server update statistics or add statistics?

Really, it can't.

So, the main point... if you're going to move a backup to another server for read-only access –you should automate some basic optimizations before setting it to RO.

SQLskills Immersion Event

IE1: Internals and Performance

Module 11: Indexing Strategies

Kimberly L. Tripp
Kimberly@SQLskills.com



fn, ln

fn like K%

and LN = 'Sm'

Katie S

Katie U

Kiera R

Kiera T

K

Indexing for AND

If the columns are doing range-based searching then the order of the secondary columns of the index might not be relevant (or even required in the key).

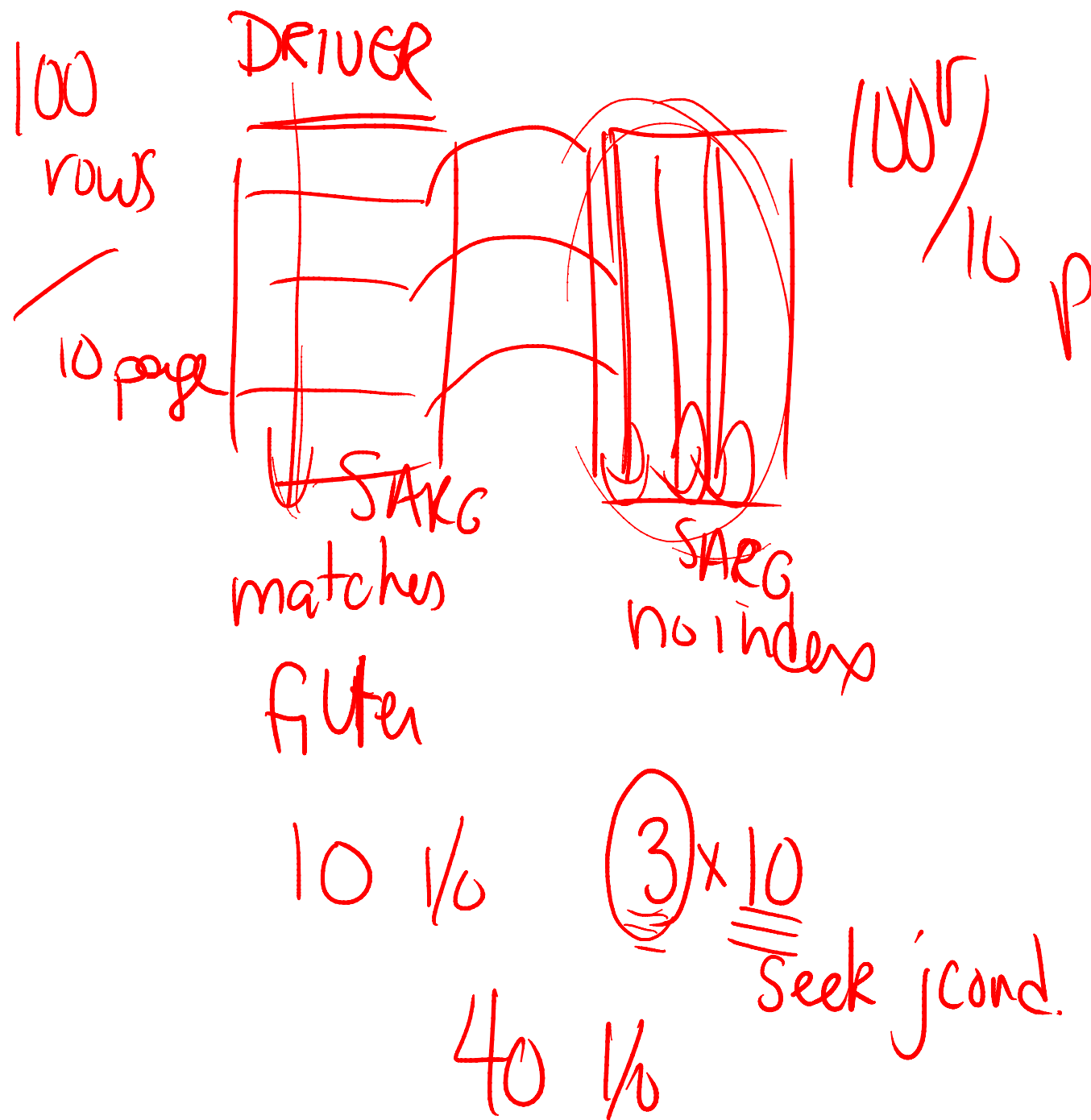
This case was shown around the conditions:

WHERE firstname LIKE 'k%'

AND region_no > 6

AND member_no < 5000

Really, it doesn't matter what follows firstname because the entire set of firstnames will need to be scanned.



S20

Here we started to talk about how LOOP joins are an iterative process. The driver (the outer/first table) is typically chosen because it has the most selective set. Any of the tables that have a highly selective search argument are more likely to be chosen as the driver. An index that aids in efficiently finding those rows is REALLY helpful!

Cost can be calculated as:

Number of IOs required for first table +

Number of resulting rows in first table * the cost for each lookup (ideally with an index on the join condition)

Loop Joins continued

If our query was:

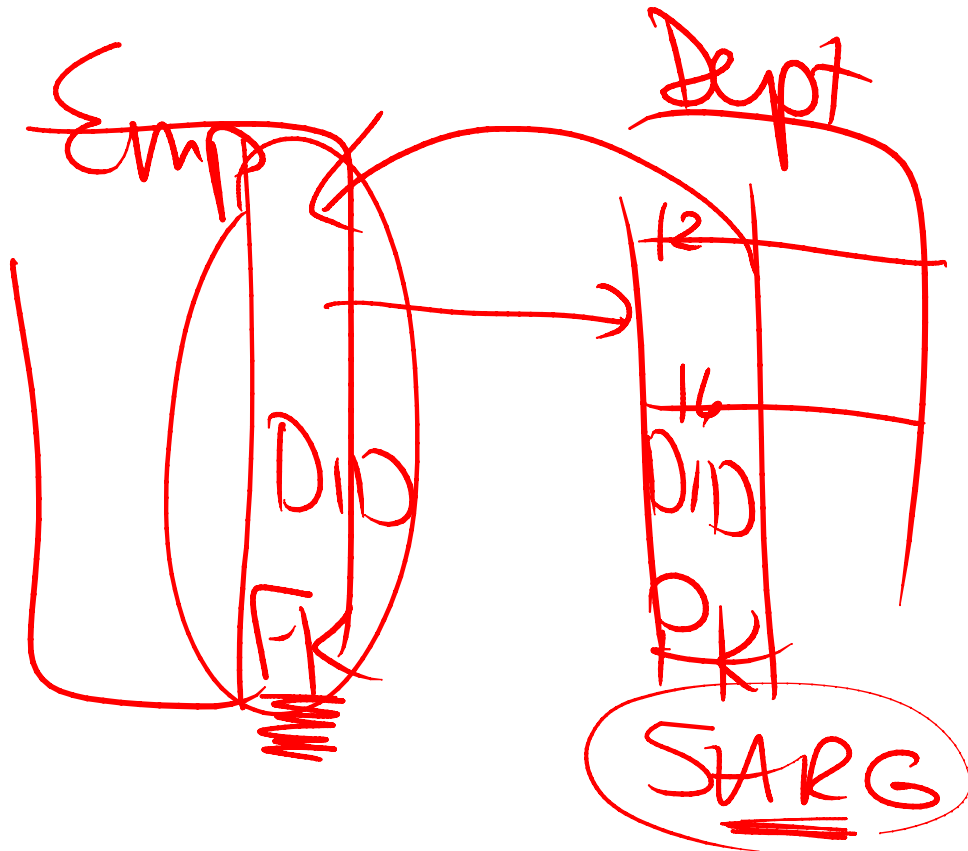
```
SELECT columns (irrelevant here)
FROM Employee AS E
      JOIN Department AS D
            ON E.DID = D.DID
WHERE D.city = 'Redmond'
```

And, there were only 1 or 2 rows for Redmond in the Departments table then that's more likely to be positioned as the driver in the join.

Once we know the DIDs then we're going to need to find all of the employees IN that department. We'll need to search against the DID column in the Employee table.

Key question:

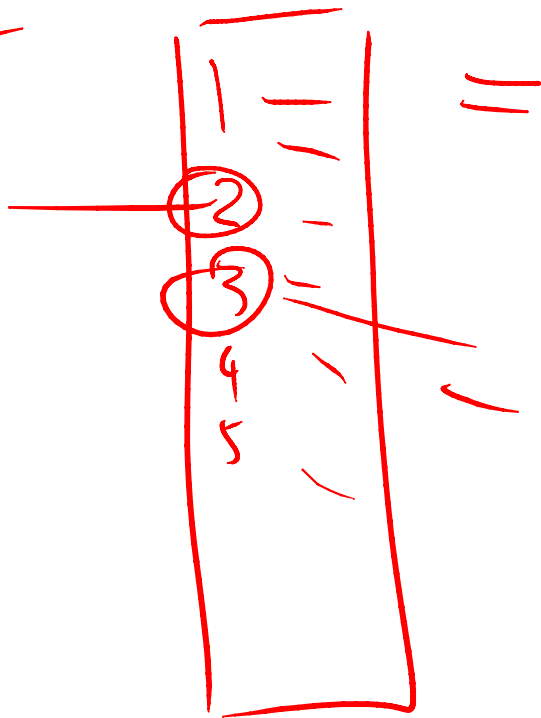
Do you have an index on the foreign key column?



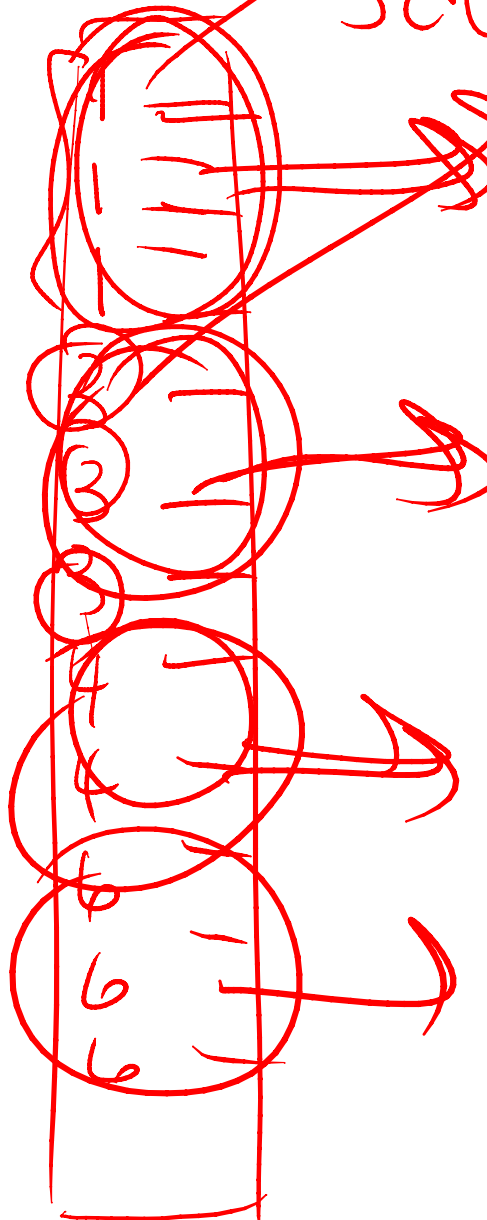
where d.city = 'Redmond'

Merge

Cost



Sales



Merge Joins

Merge joins leverage "suitably sorted sets."

More specifically, merge leverages indexes whose leading keys are on the same column.

Key question:

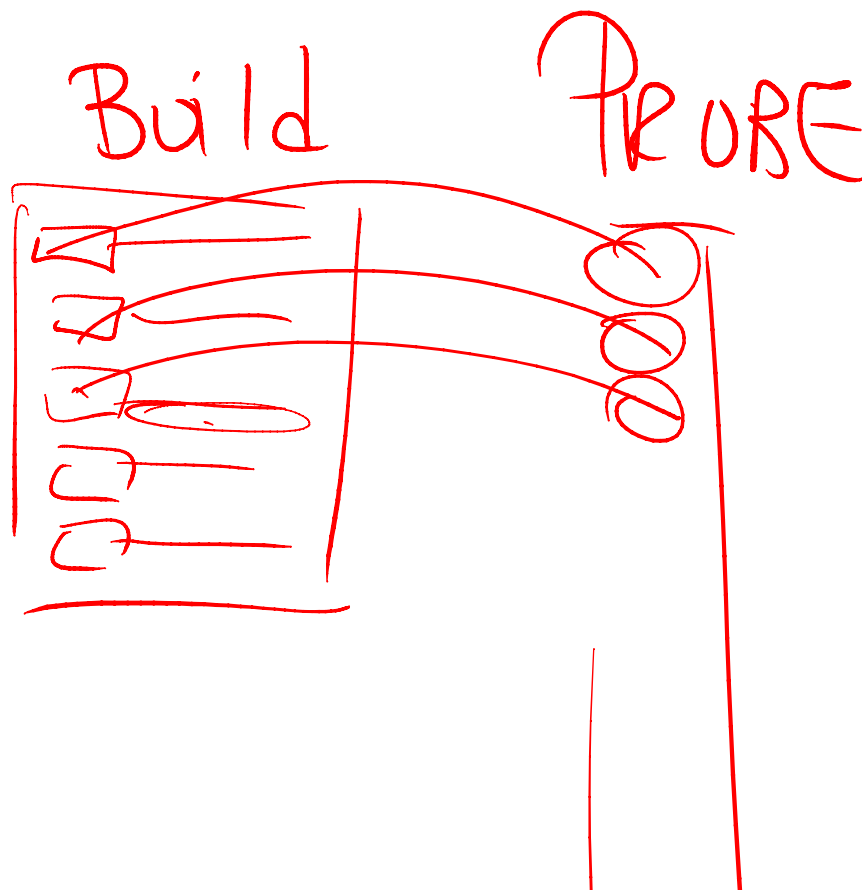
What columns do these two tables have in common?

The join column...

Corollary question:

Do you have indexes on EACH of the columns (in each table)? The one that's often missing: the join column that's the foreign key.

HASH



Hash Joins

Hash are a little more complicated. There are multiple hash types available in SQL Server and each provide different benefits. The general purpose of a hash join is to significantly reduce the number of rows that have to be processed.

More specifically, there are two phases:

BUILD phase

PROBE phase

The build phase is used to create a small structure into which the larger set can probe to determine if there's the possibility of a matching row.

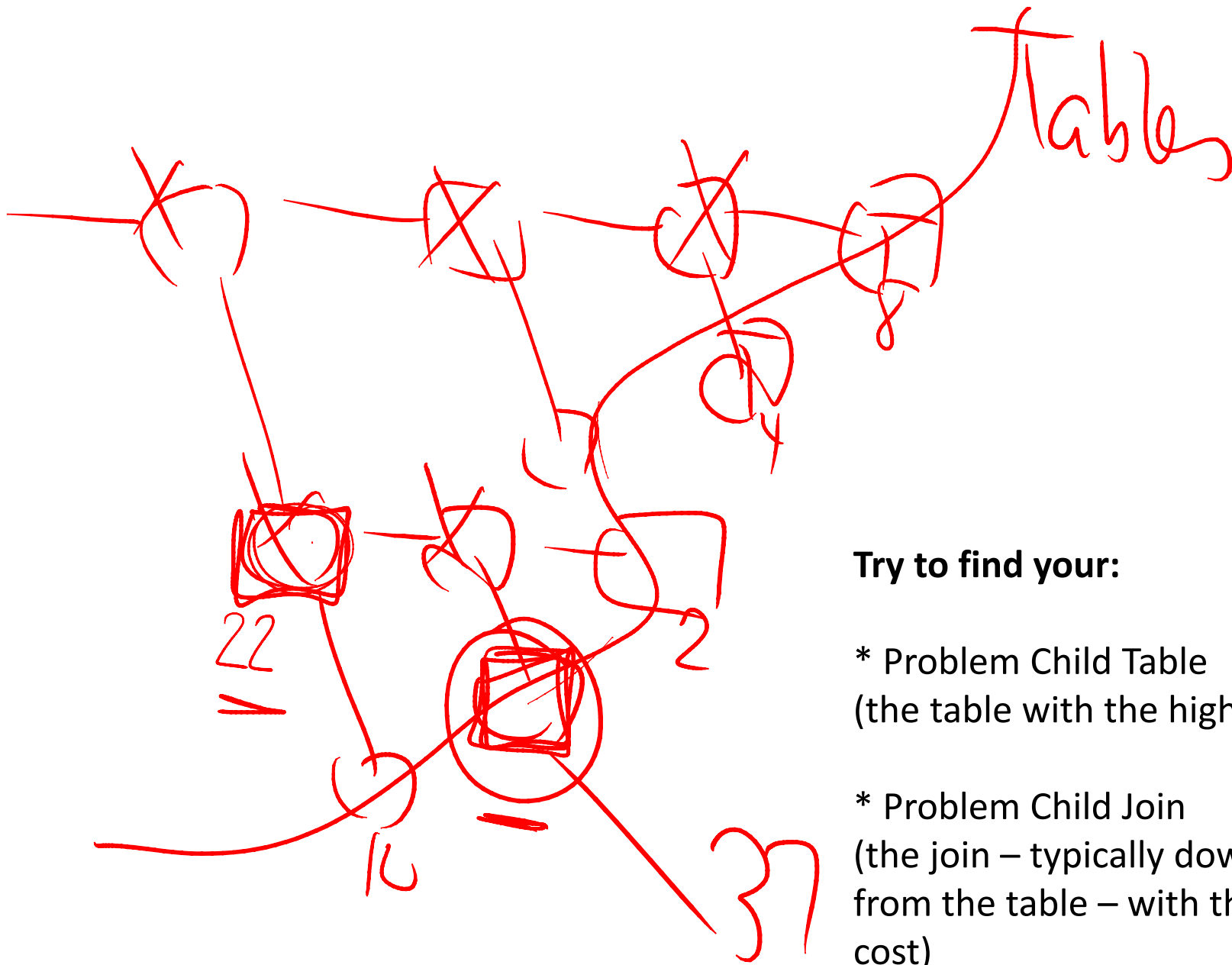
Two very good resources:

[Hash joins and hash teams in Microsoft SQL Server](#)

by Goetz Graefe, Ross Bunker, Shaun Shaun Cooper

[Query Evaluation Techniques for Large Databases](#)

by Goetz Graefe



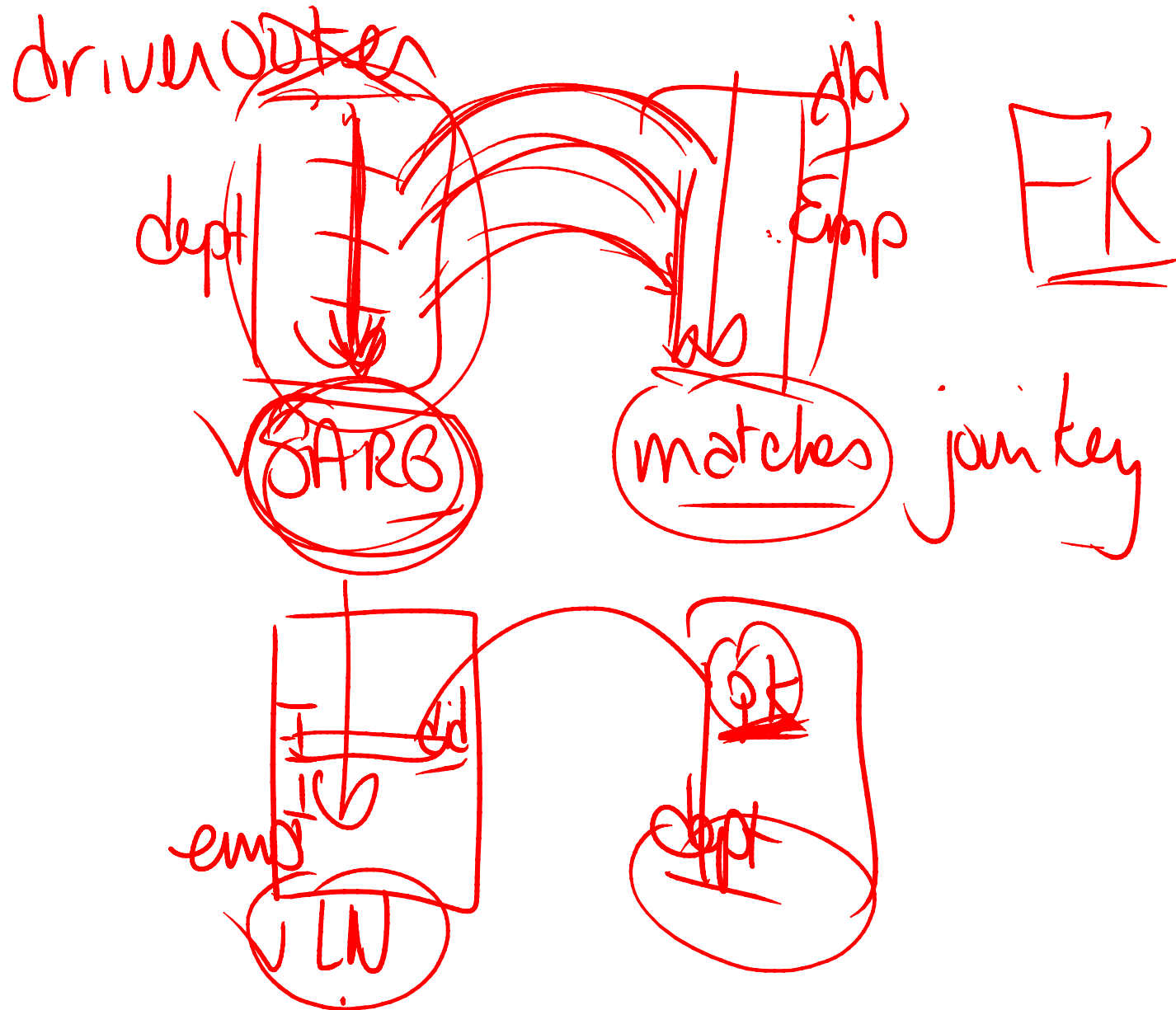
Try to find your:

* Problem Child Table
(the table with the highest cost)

* Problem Child Join
(the join – typically downstream
from the table – with the highest
cost)

S17

Tuning a large complex join takes breaking it down into smaller chunks. The things that you consider are the costs of the tables (the outer most events) and the costs of the joins. And, typically, the most expensive join is downstream from the most expensive table.



S20

An index on a join column (the foreign key) helps the performance for the referential integrity as well as *some* joins. A narrow index on the join column helps but is often superseded by wider indexes; it's still a good start. (hence the phase – phase 1)

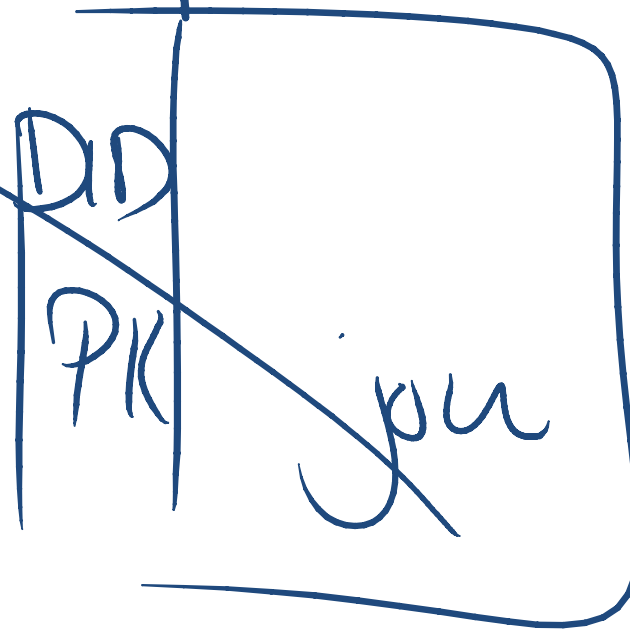
This is the most ideal for (but not limited to) a loop join.

Emp



where d.city = 'Redmond'

Dept.



Changing Join Order based on SARGs

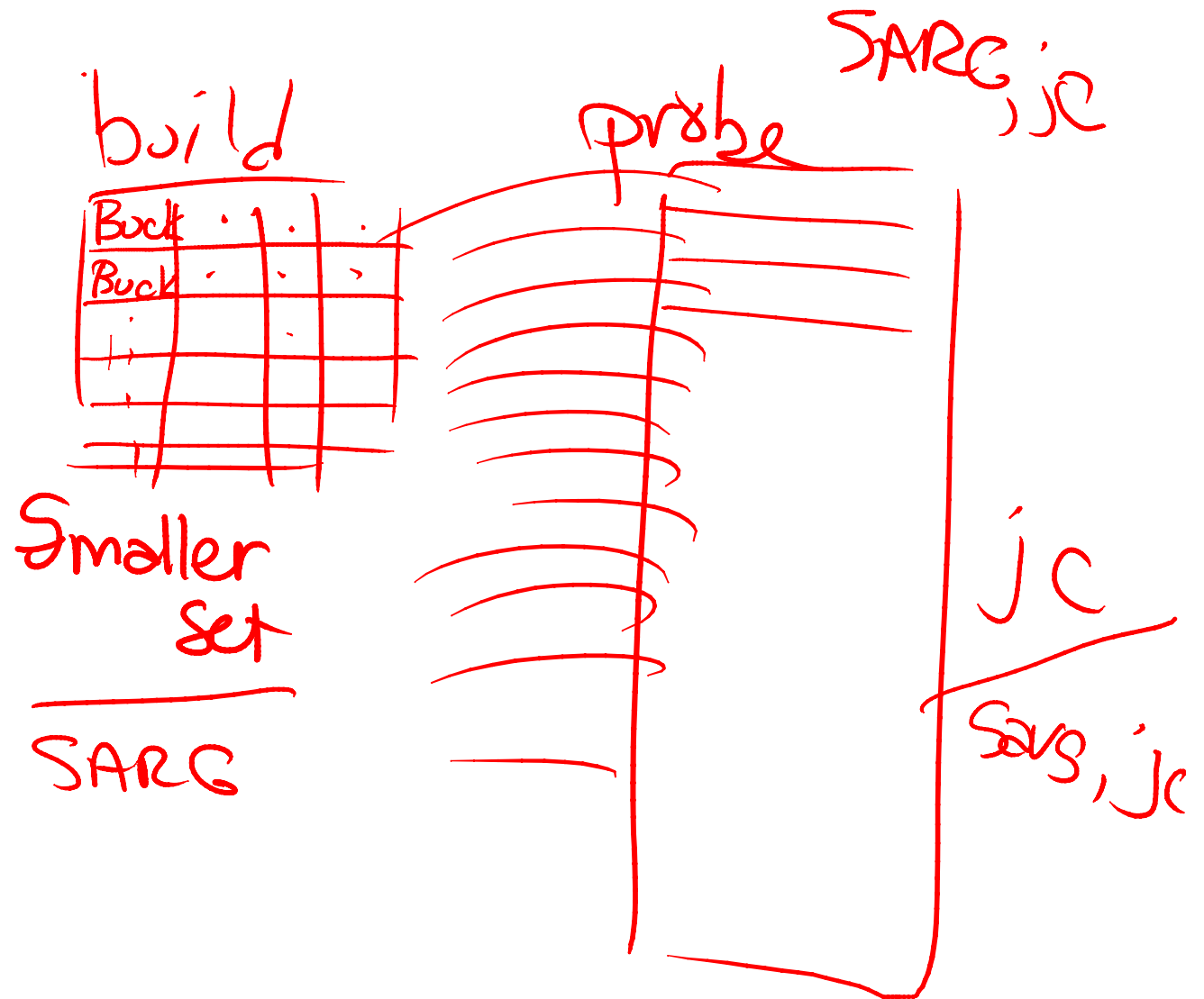
An index on a join column (the foreign key) helps the performance for the referential integrity as well as some joins because your more selective criteria might be on the referencing table rather than the referenced.

Two phase

stop

& go

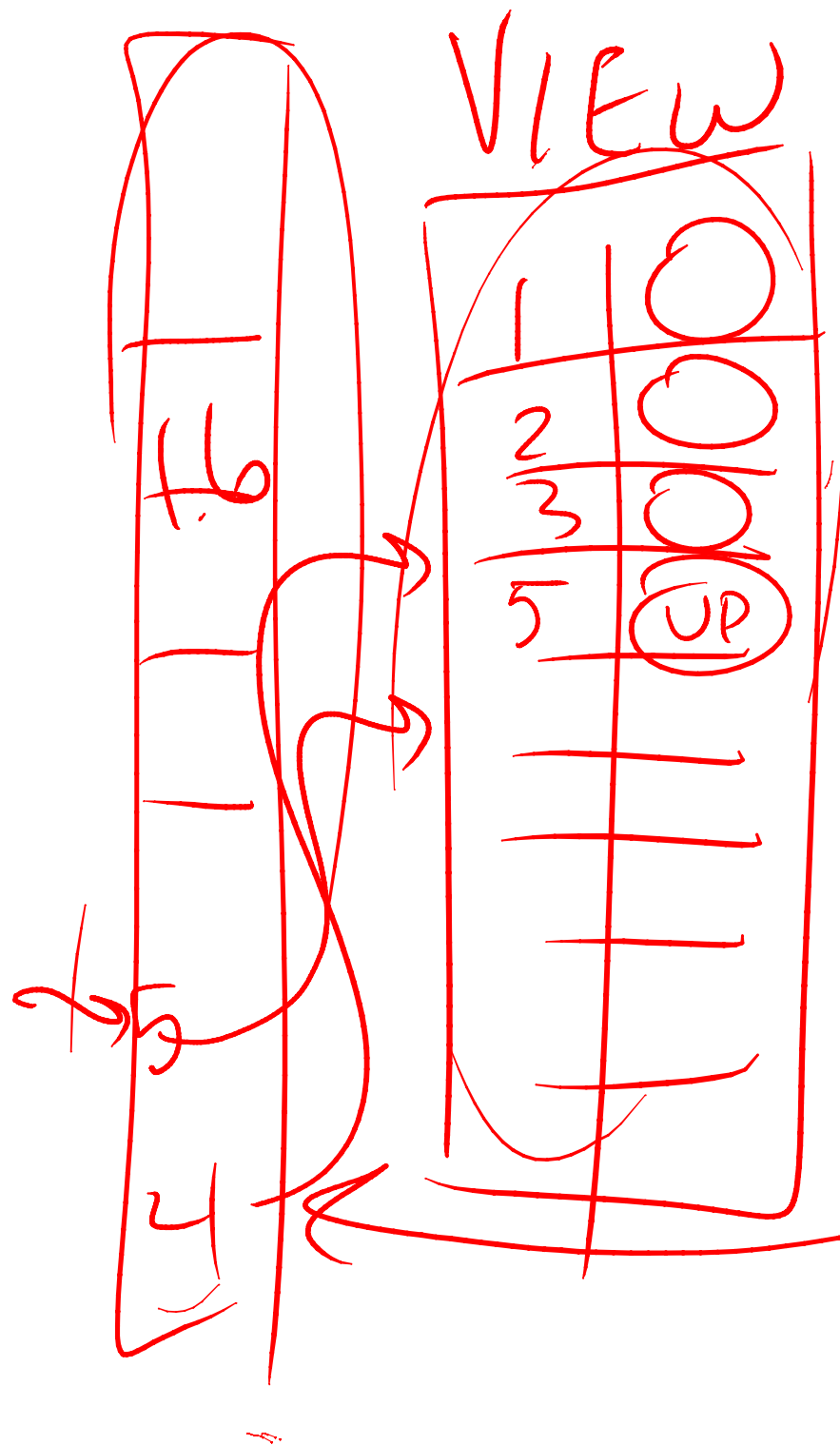
build/probe



Hash Joins are “stop and go” operations (or two-phase operations)

Phase I – Build (this is the phase where they build a table to fit in cache)

Phase II – Probe (this is where the larger set uses the join condition to “probe” into the build/temp table to output matches). Probe cannot “go” until after build completes (hence the term – stop and go).

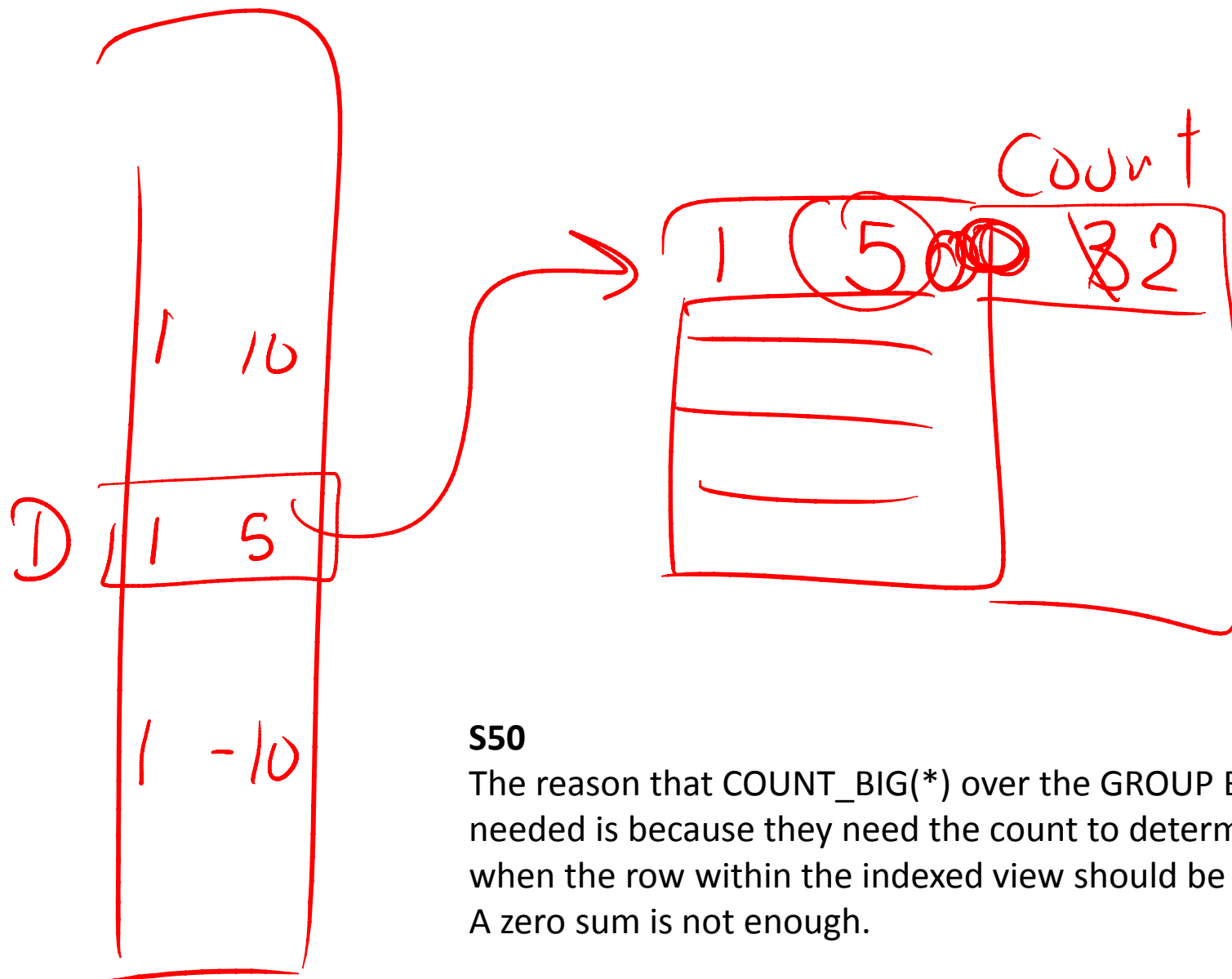


CRUCIAL IND (GRB)
ON VIEW



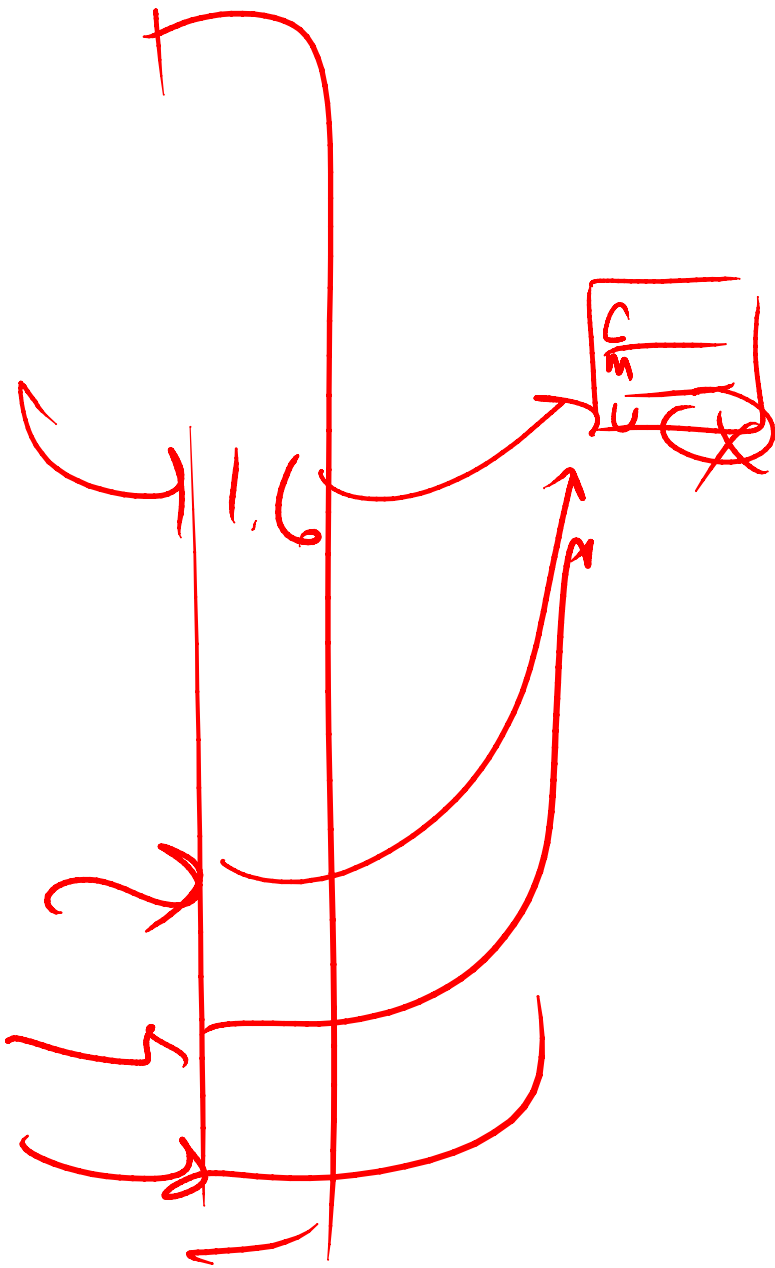
Indexed Views

Are results sets defined by a view and materialized into the leaf level of the UNIQUE CLUSTERED INDEX that's defined on the view.



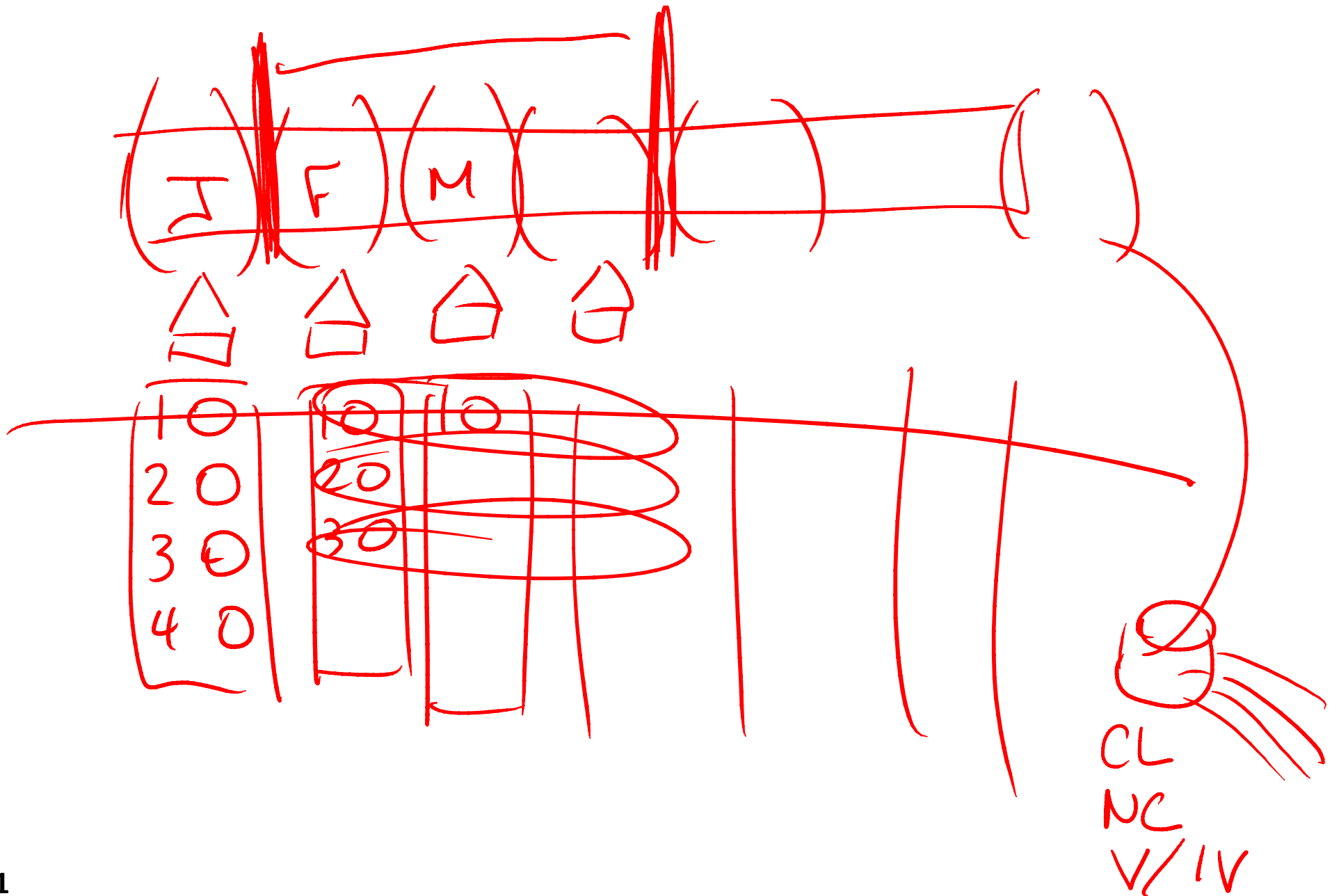
S50

The reason that COUNT_BIG(*) over the GROUP BY class is needed is because they need the count to determine when the row within the indexed view should be deleted. A zero sum is not enough.



M11

If your aggregate is too small then you can have a HOT ROW problem where all modifications are blocked trying to write to the aggregate. You'll serialize your inserts by country here... CAN, MEX, USA are the only countries with whom you do business – all US rows will have to wait as each updates the sum. This will become a terrible bottleneck.



M11

If you create a partition aligned indexed view and then request something like the `sum(sales)` for customer #1 then SQL Server can aggregate the aggregates!

On SQL Server 2008, when you're switching in – you must defined the Views and IVs in order to successfully switch in.

SQLskills Immersion Training

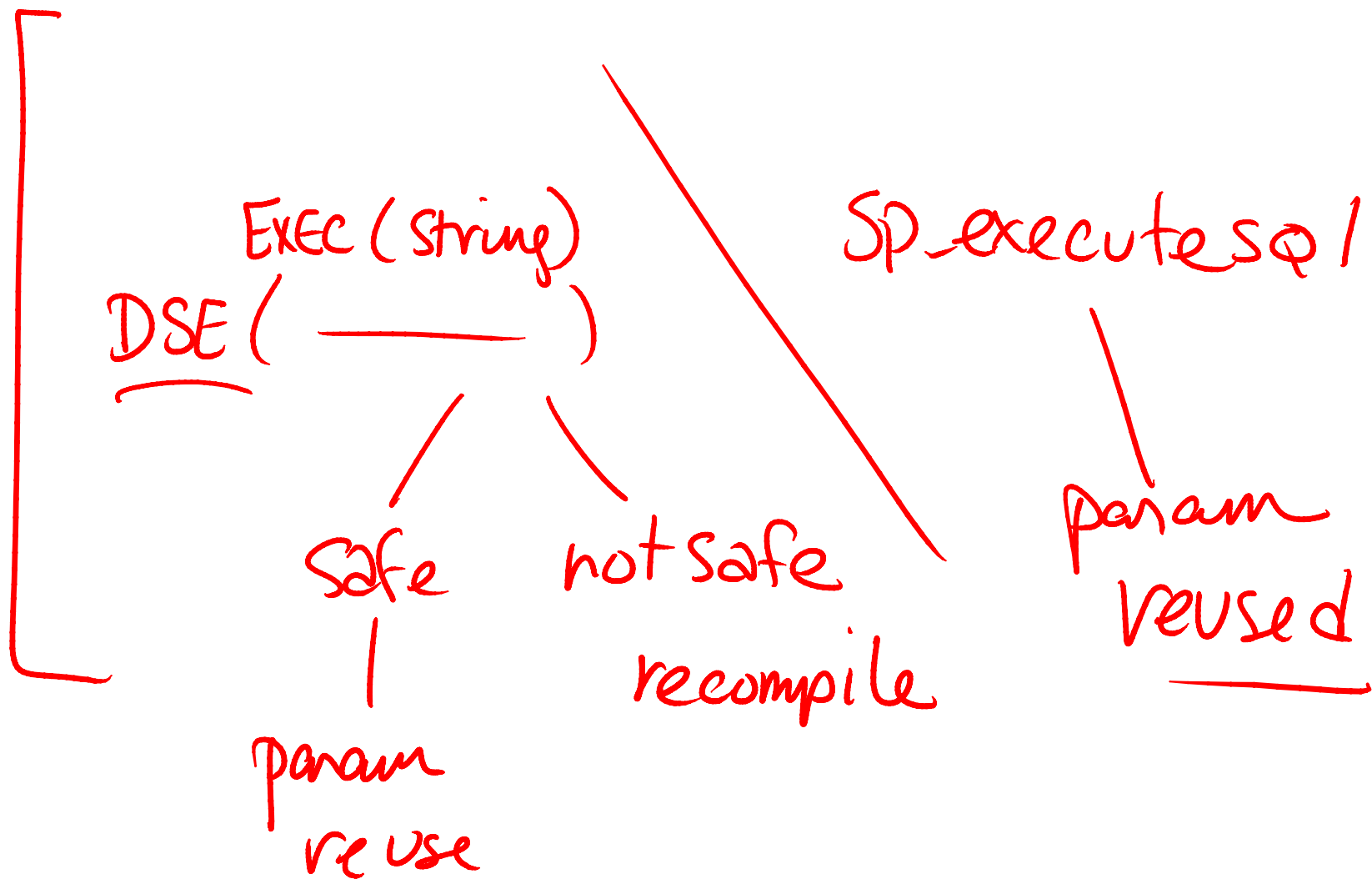
IE2: Performance Tuning

Plan Cache & Index Analysis

Kimberly L. Tripp

Kimberly@SQLskills.com



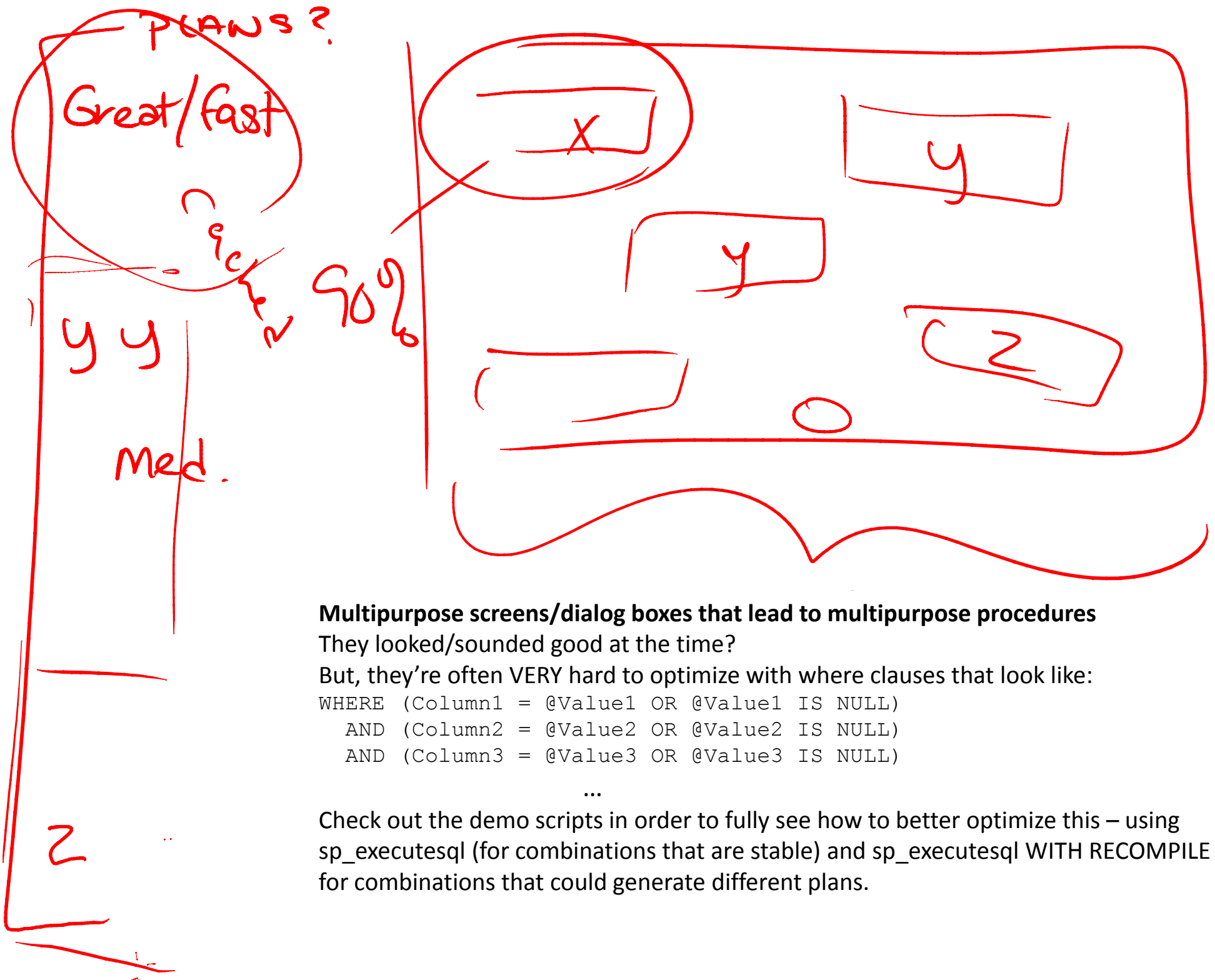


Using EXEC (string) [DSE] and/or sp_ExecuteSql [ES] inside of stored procedures

DSE and ES are not the same.

DSE is unknown until runtime. At that point the statement is treated as though it's adhoc. Because most statements are **not** going to be safe, the statement will end up being recompiled (and optimized) for each execution.

ES is forced parameterization. When a statement is stable you can use this to reduce compilation/CPU costs.



Multipurpose screens/dialog boxes that lead to multipurpose procedures

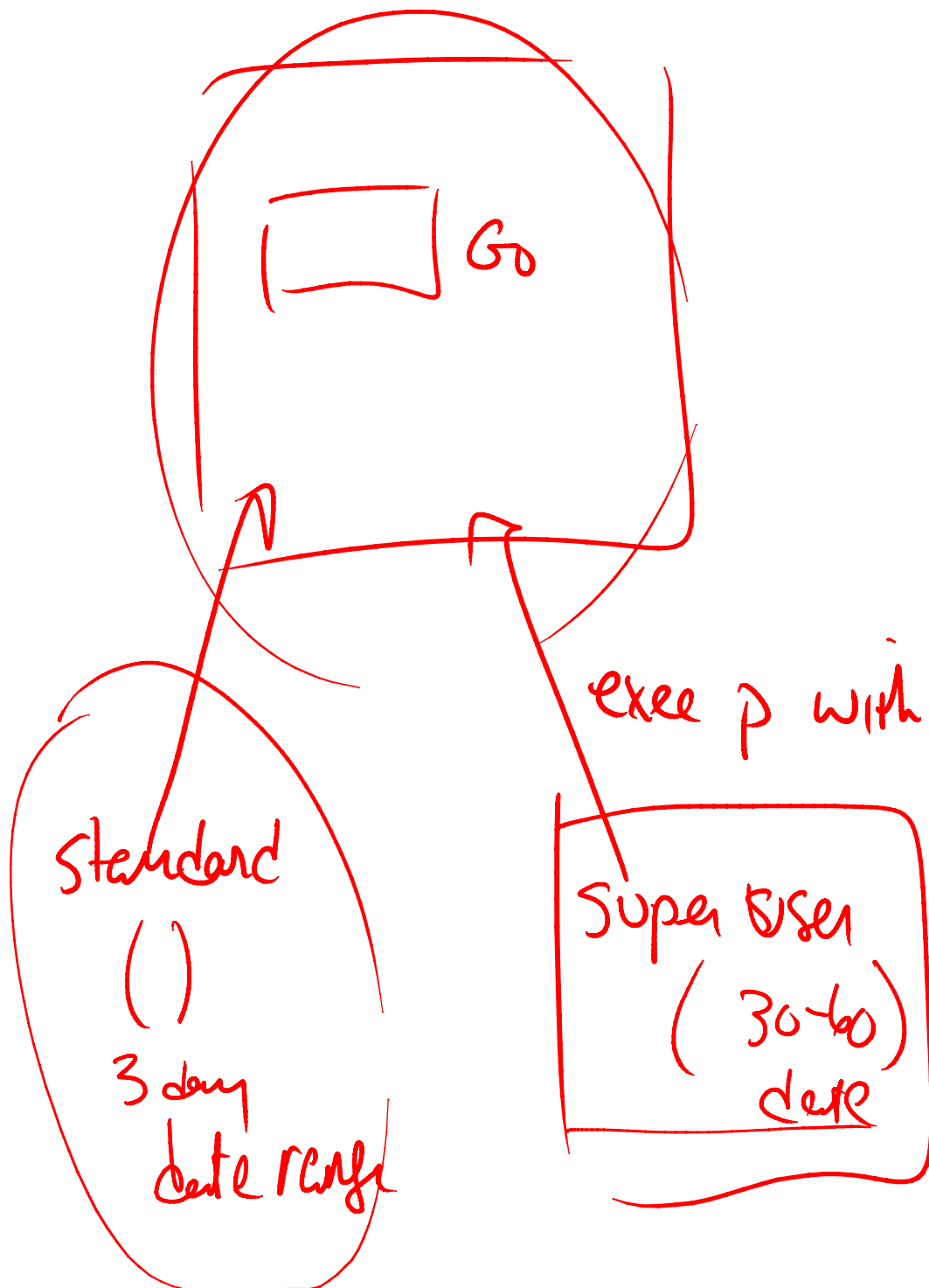
They looked/sounded good at the time?

But, they're often VERY hard to optimize with where clauses that look like:

```
WHERE (Column1 = @Value1 OR @Value1 IS NULL)
      AND (Column2 = @Value2 OR @Value2 IS NULL)
      AND (Column3 = @Value3 OR @Value3 IS NULL)
```

...

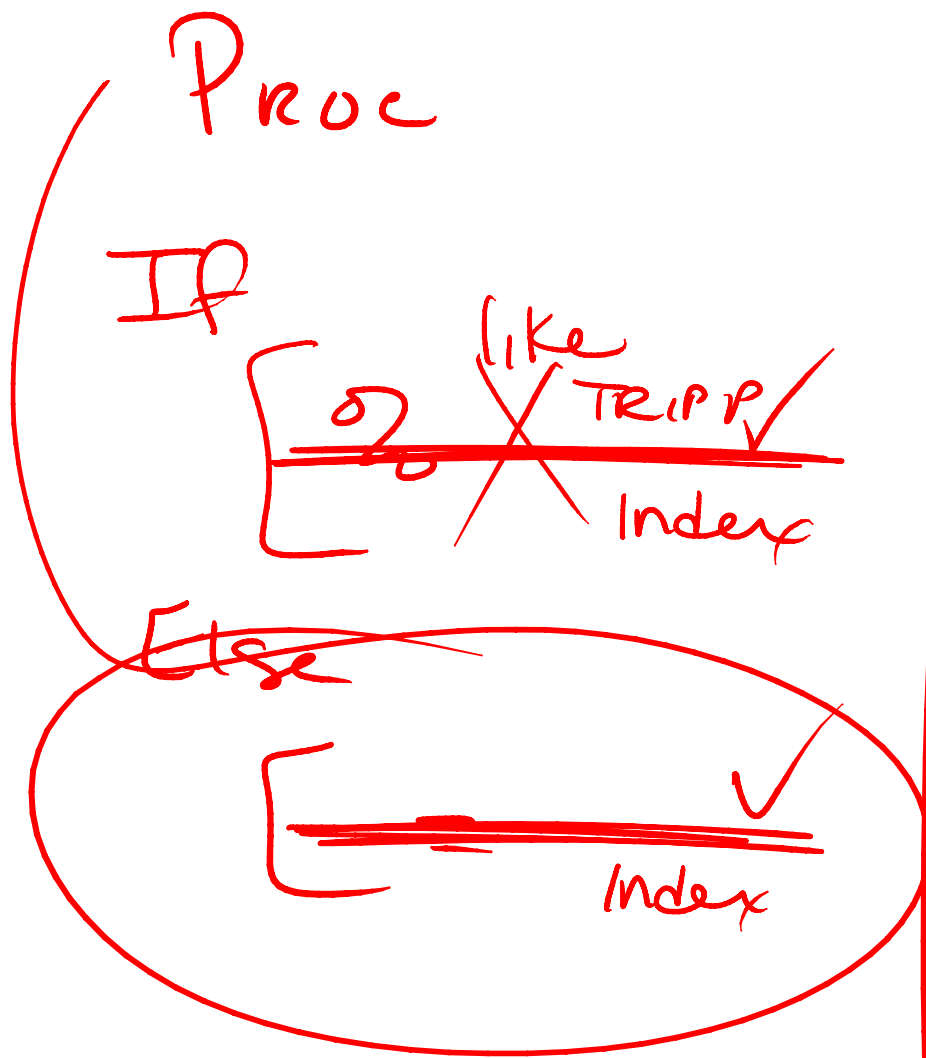
Check out the demo scripts in order to fully see how to better optimize this – using `sp_executesql` (for combinations that are stable) and `sp_executesql WITH RECOMPILE` for combinations that could generate different plans.



Options for different executions

If we largely execute with “standard” users and we absolutely want THAT plan to be in cache and to be re-used then we could:

- 1) Use option (optimize FOR...) and use a literal that will generate a plan that’s great for the “typical” scenario. However, the performance won’t be great for the atypical (super user) scenario.
- 2) We could use EXEC for the typical scenario and EXEC proc WITH RECOMPILE for the super user scenario. The benefit of this is:
 - 1) We get a cached plan for the typical user (no CPU/plan stability)
 - 2) We get a specific plan for the atypical user (the super user) as opposed to a possibly inefficient plan (so, we have to compile but we’re only compiling this one type of plan)



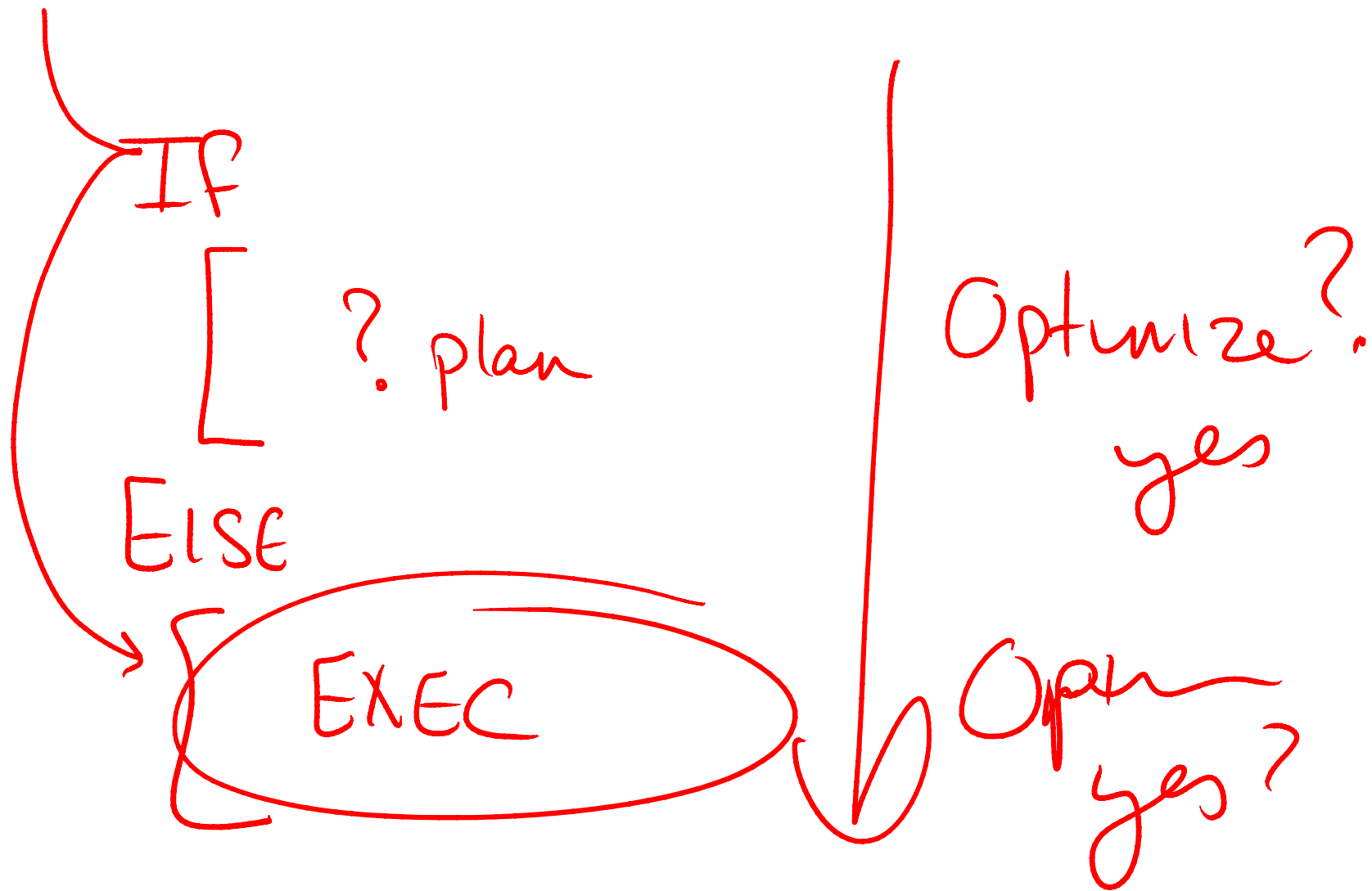
Proc

_____ ✓

_____ ✓

Conditional logic and modularization

This is another way of looking at the same thing from the prior diagram. Really, don't think of the "logic" in your procedure; think only of the procedure as a list of statements (regardless of conditional logic). SQL Server tries to optimize ANY statement that's "*optimizable*" with the parameters that were supplied (regardless of those specific parameters apply for *that* execution). This is another case where you can end up with an inefficient plan because of parameter sniffing.



Conditional logic and modularization

Creating branching logic for different types of parameters seems logical but because SQL Server optimizes the process of optimization (where they try to optimize anything that's optimizable), you can often end up with a plan that's not ideal.

You are much better off breaking that code into smaller subprocedures – SQL Server will never step into a subprocedure unless it's executed and in this case it will only be executed with exactly the right parameters.

CR PROC —
(param1
 pa 2
 —)

} Sniffed

as

Declare var — unknown why?
 we don't know
 until EXECUTION

Select -- -- param?

Select -- -- param?

Parameters are “sniffed”

SQL Server uses parameters to optimize. Sniffing implies that SQL Server uses the histogram to evaluate the data selectivity. For the FIRST execution parameter sniffing is great. It's the subsequent executions that can suffer from parameter sniffing (and therefore end up with PSP = parameter sniffing problems).

Variables are “unknown”

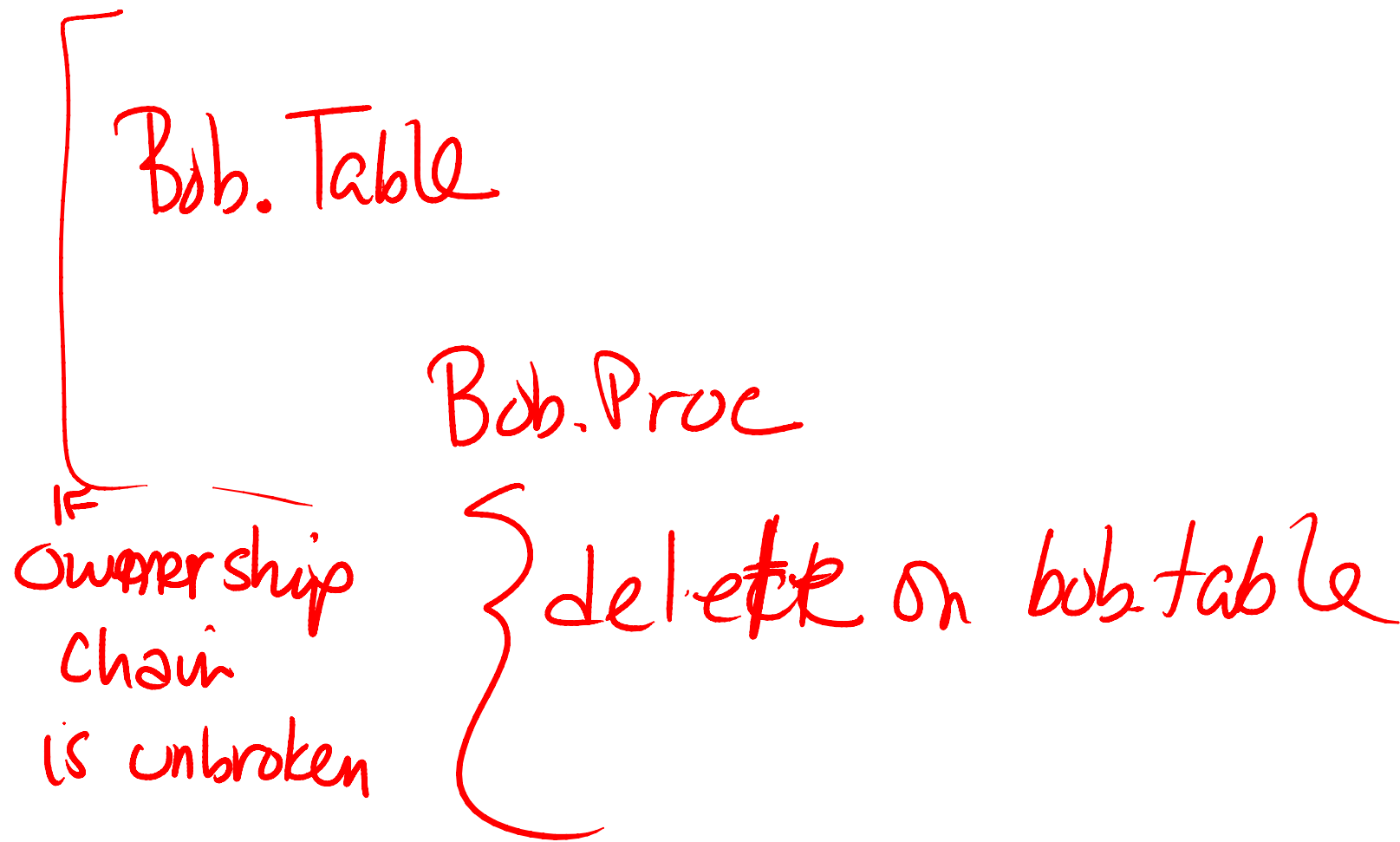
Variables are only known at runtime as the statements execute and the variables are assigned. As a result, they are unknown during optimization. With an actual value, SQL Server cannot use the histogram. Instead SQL Server uses the density vector for the “average” numbers of rows that meet that criteria.

DEFAULT
Proc

DSE — protects against SQLinj
execute under context
CALLER

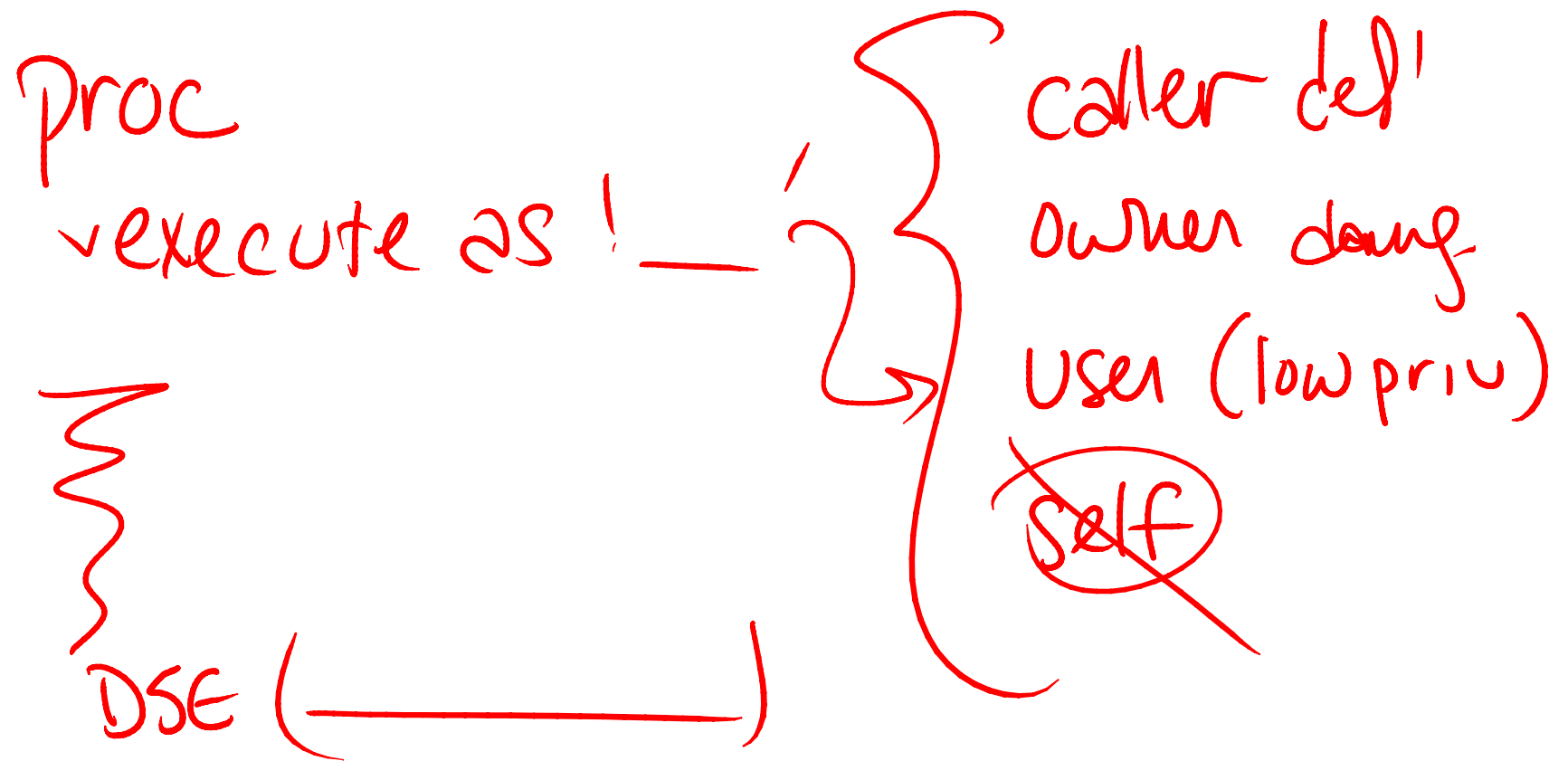
Dynamic String Execution and SQL Injection

TO prevent SQL Injection, SQL Server requires that strings execute under the context of the caller.



Without Dynamic String Execution permissions flow through the ownership chain

The idea is that when the ownership chain is unbroken then direct permissions (for example, to do a delete) are not necessary. As an owner you are giving rights to a process; this procedure can be protected with more business rules/logic and even reduced/isolated to specific rows. The end user does NOT need direct delete permissions to be able to execute the procedure as long as they've been given execute rights on the sproc.



Dynamic String Execution, EXECUTE AS and SQL Injection

People complained that granting permissions directly to the user was a problem... so, they wanted an alternative. Enter: EXECUTE AS.

While it solves the problem of permissions, it adds more potential for SQLInjection.

So..... Check out these posts:

[Little Bobby Tables, SQL Injection and EXECUTE AS](#)

["EXECUTE AS" and an important update your DDL Triggers \(for auditing or prevention\)](#)

✓ fn .
✓ (fn + ln) ?

Stable plans vs. unstable plans – how do you know?

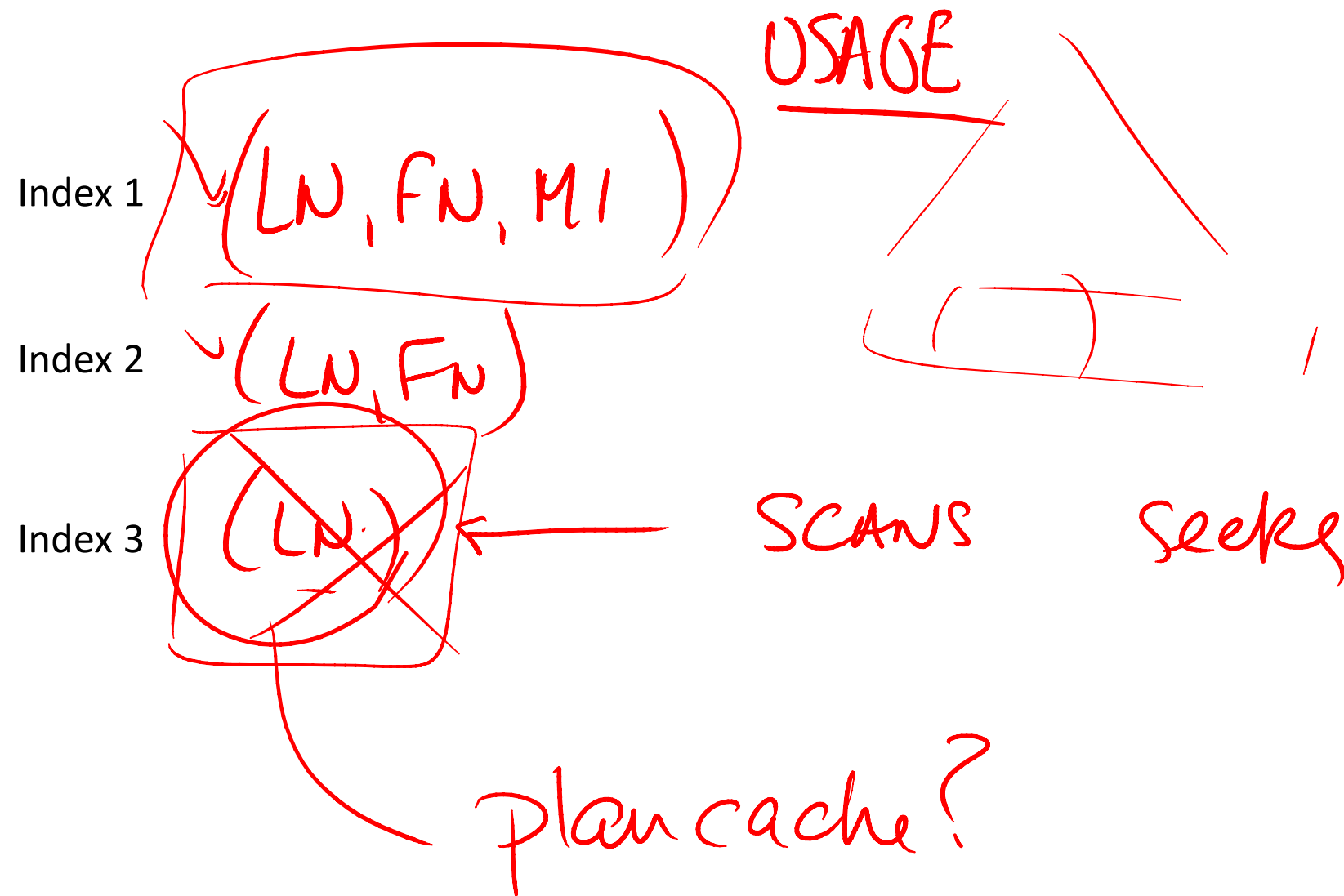
- You could test different combinations using execute with recompile.
- You might know because of selectivity?

* ~~fn + ln + mno~~ ← stable/cons
✓ ln

* ~~ln + mno~~ ←
* ~~mno~~ ←
* ~~mno + fn~~ ←
} Should not
ree.

S38 – options that better balancing recompiling too much vs. not recompiling enough

This is from the multipurpose procedure demo. There are 3 parameters and 7 possible combinations that can be submitted. Instead of adding `OPTION (RECOMPILE)` to the statement (which doesn't always work [remember, it has a very checkered past]) you can build the statement dynamically and then programmatically (by setting a `RECOMPILE` flag to 0 or 1) decide whether or not `OPTION(RECOMPILE)` should be added to the dynamically constructed statement. Then, you can use `sp_executesql` to place ALL seven statements in cache. Those without the `OPTION(RECOMPILE)` clause will be cached. Those with – will get a plan on each execution.



Should you drop redundant indexes? It depends. 😊

It's TRUE that query that uses a left-based subset of one index (let's say that a query uses index 2) that is COULD use a wider version of the same (left-based) index. It could use Index 1 instead of 2.

The problem is that there could be specific uses that warrant the smaller version. Questions to ask: Is the smaller version used by scans? Are they executed VERY frequently? How important are they?

Key Inc

$C1 = C2$ ~~$C3$~~

$C1, C2, C3$ no inc

Key = navigation

Key must be pres.

Key Inc

<u>LN, FN</u>	
<u>LN</u>	
<u>LN, FN, MI</u>	
LN	SSN
//	
<u>LN, FN, MI</u>	(^{inc} SSN)

Consolidation?

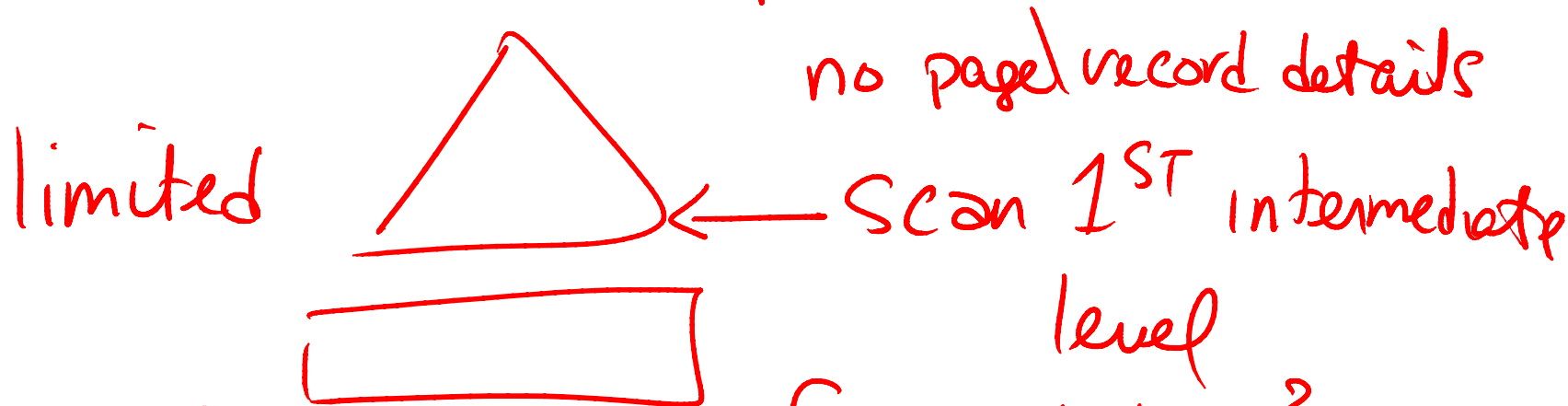
You can CONSIDER removing indexes that are left-based subsets of others (in terms of the key). However, be sure to preserve the key as that's used for navigation/seeking. The columns that are included can be combined in any order.



Index Health – Checking for fragmentation

For a limited mode query against `sys.dm_db_index_physical_stats`, SQL Server uses a scan of the first intermediate level (index level 1) to analyze for the value: `avg_fragmentation_in_percent`. If your analysis/rebuild scripts use this to determine fragmentation than you'd be better off just doing a limited scan. What `SAMPLED` and `DETAILED` offer are page-specific details (page density, forwarding pointers, etc.) and if you're NOT using those (most people aren't using these during their automated maintenance process). So, don't waste time with a `SAMPLED` or `DETAILED` scan during your maintenance window!

Sys. dm-db-index-physical-stats



Good for avg-fragmentation-in-percent fragmentation?

Sampled } Avg page density, page fullness
detailed } row size

→ 1/100 of leaf / detailed reads all pages
all levels

FN like K%

rno > 6

mno < 5000

fn, mno, rno

Katie 6 4

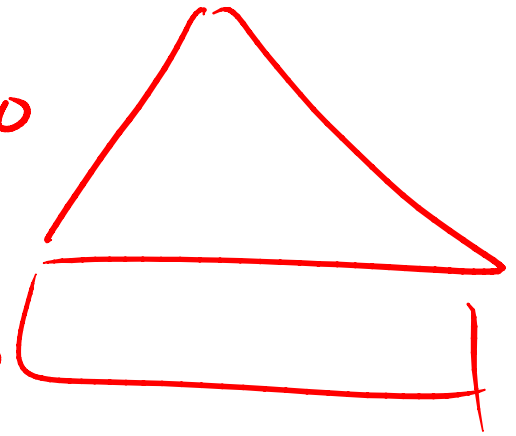
Katie 7892 6

Kiera

Kiera

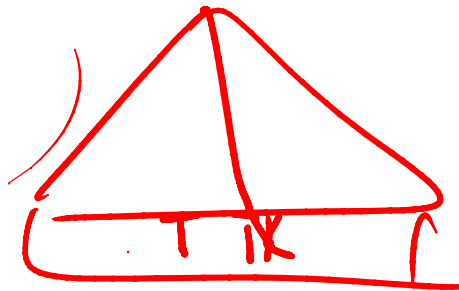
Kimberly

Kimberly

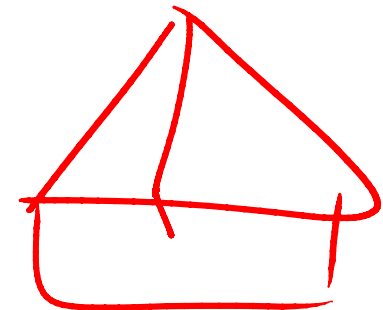


FN = Kimberly

LN = TRUPP



LN, FN



FN, LN

Adding more indexes

Here we were discussing the order of columns. If there are range-based predicates than it doesn't usually matter which column is second. You tend to want to put the most selective first (in terms of the PREDICATES supplied). If all of the predicates are equality-based then it doesn't really matter. You're best ordering them by the LEFT-based combination that's used most commonly.

KEY

Density_Vector

c1

c2

c3

c4

c2

c4

c2

c3

c3

c4

c2

c3

c4

Multi-column, column-level statistics

Only DTA recommends multi-column, column-level statistics...

If you are about to trust (and create) a “green hint” recommendation (which comes from the Missing Index DMVs) then you might want to first run the query through DTA. If you end up creating the index that DTA recommends then you should also create the recommended stats (for that same table). These statistics can be helpful to the optimizer to better understand non-left-based subsets of the index for other possible uses.

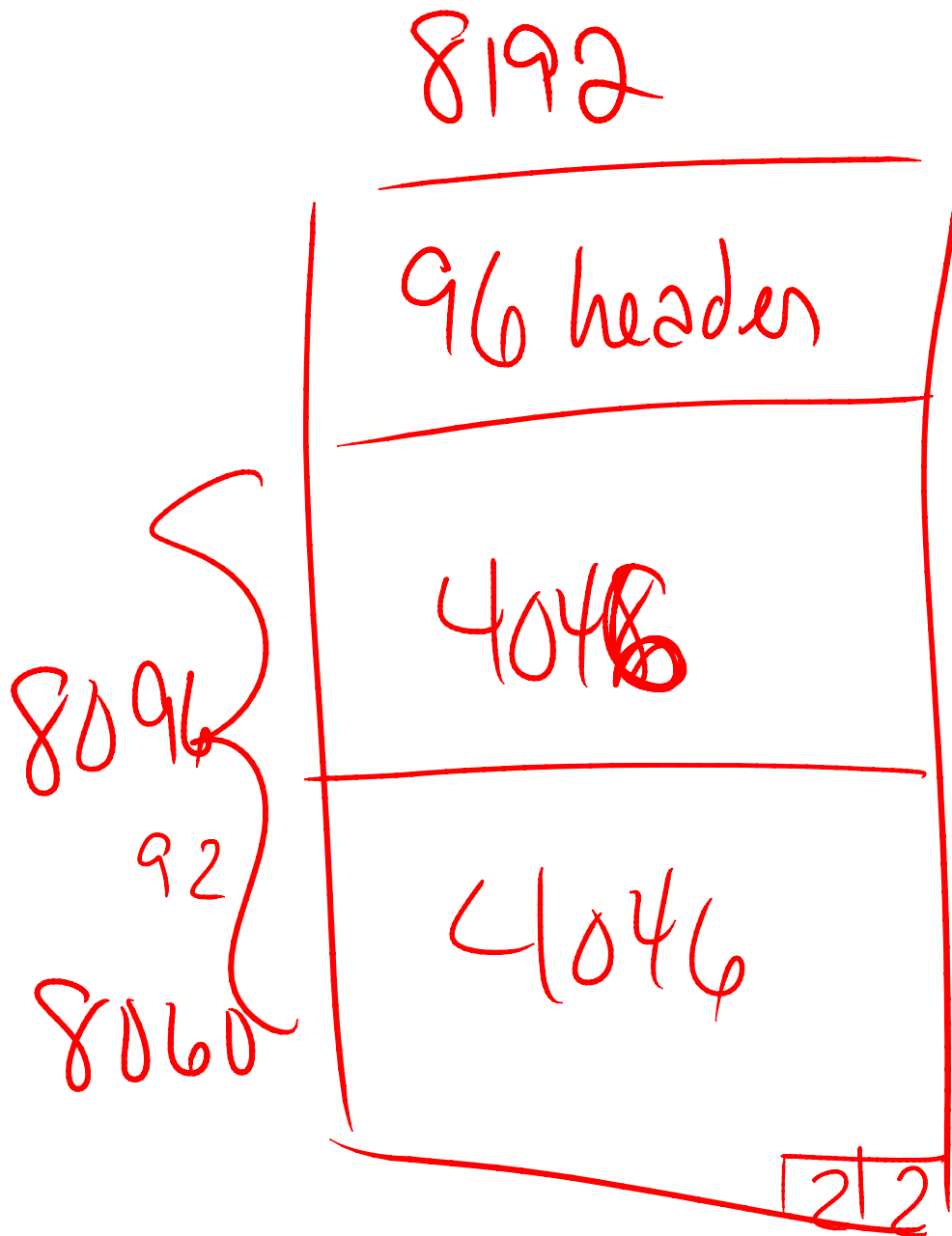
SQLskills Immersion Event

IE1: Internals and Performance

Module 6: Table Design Strategies

Kimberly L. Tripp
Kimberly@SQLskills.com





S4

This was a reminder that when you're calculating row size (as well as page density) that you can use 8096 for data rows. Every row requires 2 bytes in the slot array so two rows could each be 4046. However, a single row cannot be more than 8060. The 8060 comes from:

8192 bytes (page size)

- 96 bytes in the header

= 8096 as the maximum amount of space for data

A row must have a header of 4 bytes

= 8092 and the SQL team took "32 bytes" for future growth.

Some of which they're already using in SQL Server 2005 and higher. For row versioning, each versioned row requires a 14 byte offset. If they had allowed a row to be 8096 – where would they have put the offset?

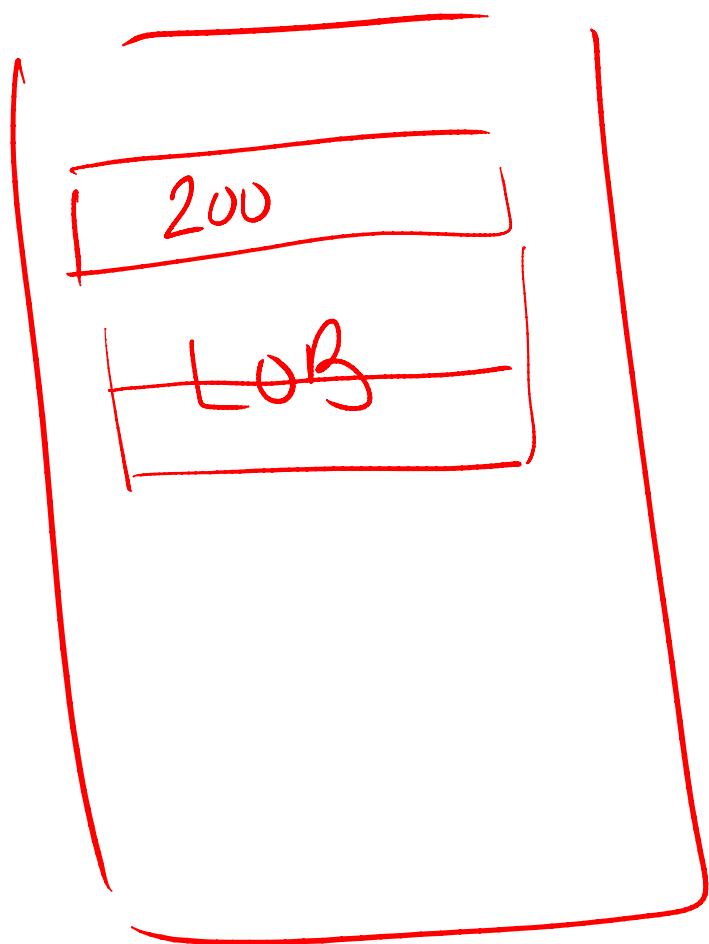
				2	2
	VC	VC	VC	VC	VC
_____	X		X	X	2
_____				X	2
_____		X		X	2

Column order does not matter... it depends!

OK, this diagram should remind you that leaving the columns that are most likely to be null – at the end of the row definition *COULD* save some space.

Blog post: Column order doesn't matter... generally, but - IT DEPENDS!

<http://www.sqlskills.com/BLOGS/KIMBERLY/post/Column-order-doesnt-matter-generally-but-IT-DEPENDS!.aspx>



LOB

100-200
bytes MB

majority

< 2000 bytes

S7

Default behavior of a *new* LOB (max or XML) is that they will be put "in_row" if they fit (if the total row size is under 8060).

If you have a lot of small LOBs (for example 2K LOBs) then all of your rows will be really wide. If you're doing a lot of scans and NOT interested in always returning the LOBs then you're going to waste a lot of pages/IOs to get the LOBs into cache when you don't really need them.

Consider pushing these small LOBs off-page instead!

S8 Implicit Conversions

When a predicate is being evaluated there are two sides to the equation:

The column expression & The variable expression

When one of the expressions has a higher data type AND they're implicitly compatible then SQL Server will need to bring the other expression UP to the same type.

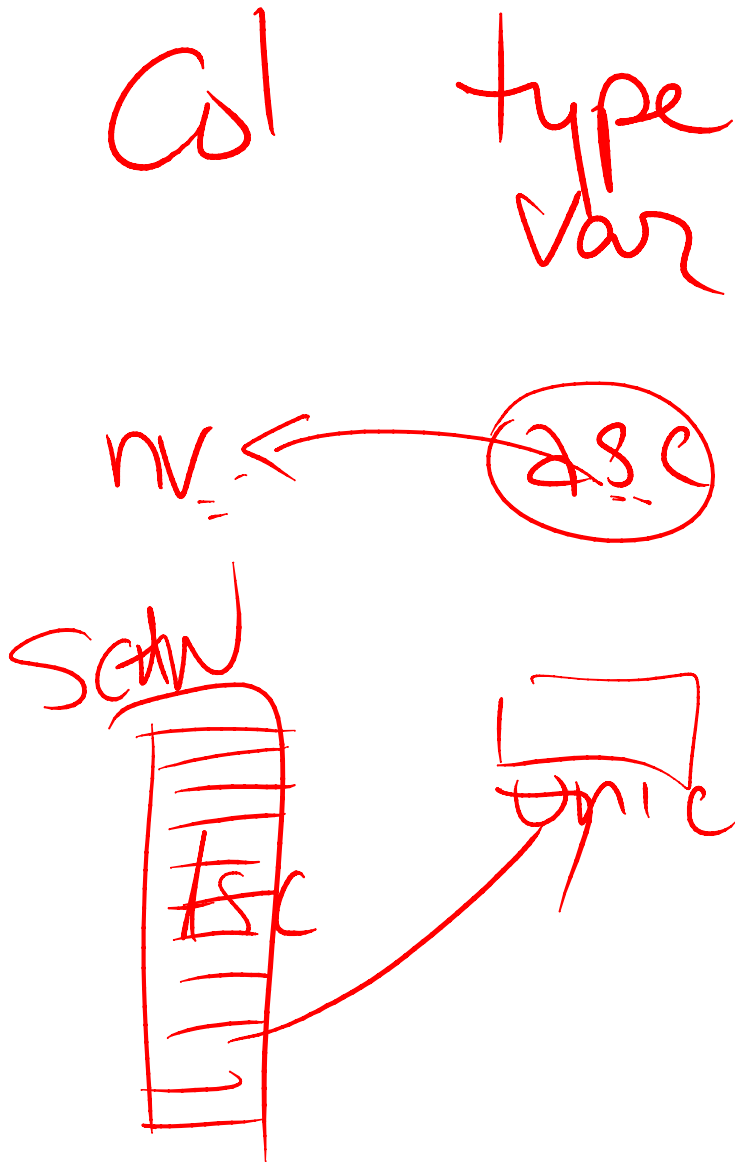
If the variable expression is the LOWER type then it's easy, only that variable has to be converted.

The problem starts when the column expression is the lower type – then the column has to be converted. This results in a scan of that column (could be an index scan but if a good index doesn't exist then it might be a table scan).

You can see implicit conversions in cache – check out Jonathan Kehayias' blog post on this:

Finding Implicit Column Conversions in the Plan Cache

http://sqlblog.com/blogs/jonathan_kehayias/archive/2010/01/08/finding-implicit-column-conversions-in-the-plan-cache.aspx

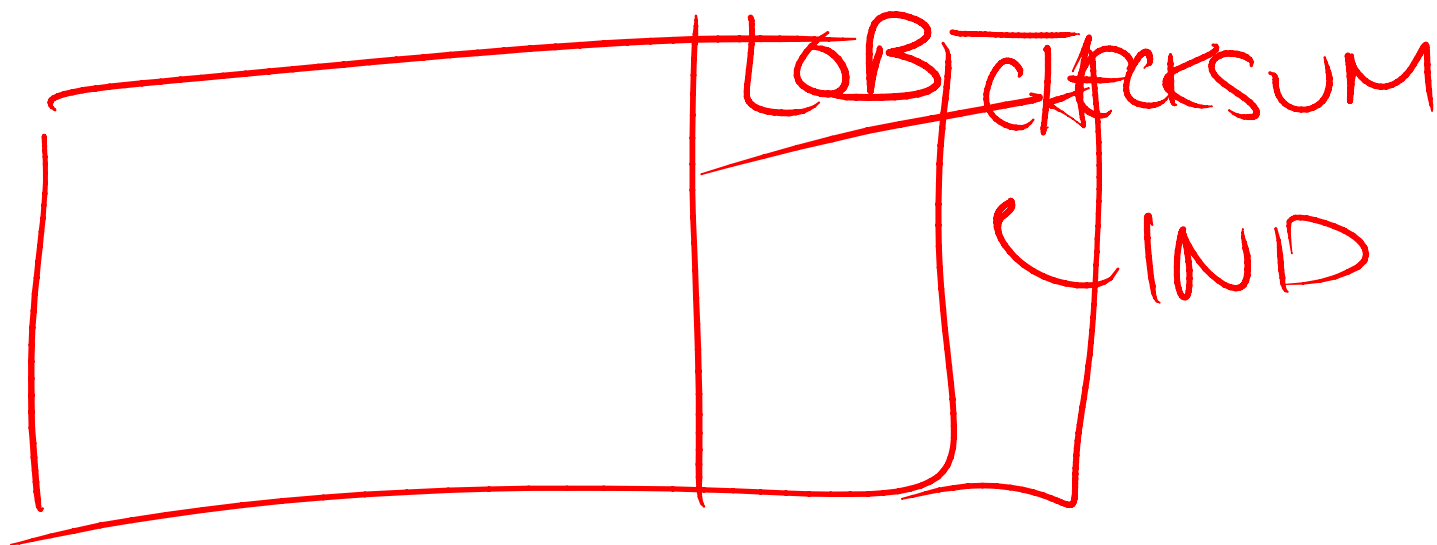


Placeholder Rows

Client $\longrightarrow$ ID def. def
VC VC VC VC

P/C
C
C
C
C





where LOB
 and check sum = checksum(LOB)

This was a tangent where we talked about a way to optimize lookups for LOB data.

One of the things that always causes people grief is that columns over 900 bytes cannot be indexed. So, if you have a WHERE clause like this:

```
WHERE lobcolumn = 'lob value.....'
```

SQL Server has to do a table scan to get all of this data.

Another trick is:

Add a CHECKSUM column that checksums a particularly large column (note: legacy LOB column types: (n)text, image AND XML are not supported – see this link: [http://msdn.microsoft.com/en-us/library/ms189788\(v=sql.105\).aspx](http://msdn.microsoft.com/en-us/library/ms189788(v=sql.105).aspx)) and then rewrite your queries to be this:

```
WHERE lobcolumn = 'lob value.....'
```

```
AND checksumcol = CHECKSUM('lob value.....')
```

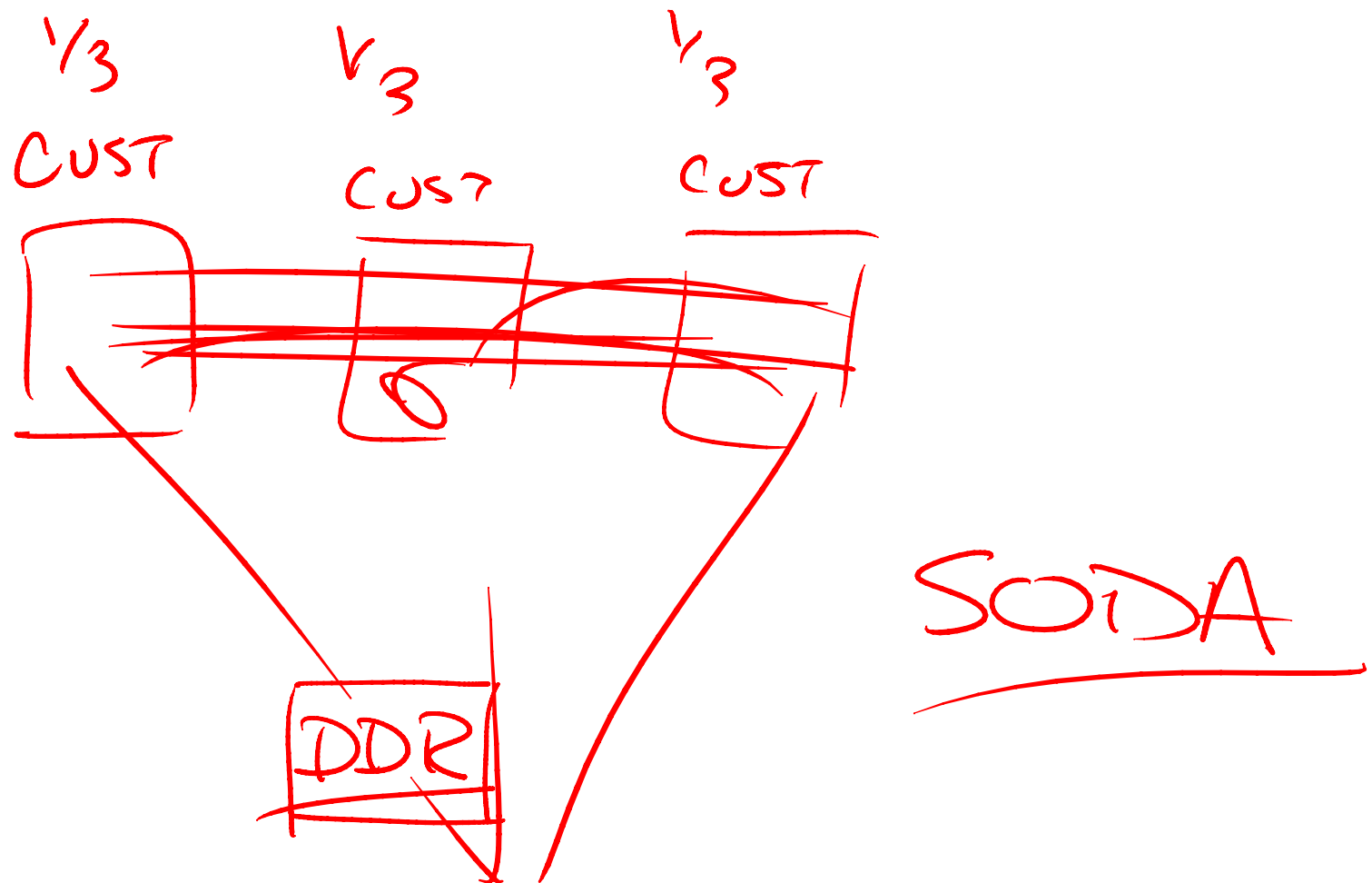
SQLskills Immersion Event

IE1: Internals and Performance

Module 12: Partitioning

Kimberly L. Tripp
Kimberly@SQLskills.com





M12

- Partitioning in our module was all within a single database. However, we had a side/note about sharding and scale-out design. The most important thing that I can highlight is that scaling out is most ideal through middle-tier DDR (data-driven routing) where the applications are directed to the appropriate server. If every user randomly goes to any of these instances and all requests go through [distributed partitioned] views then performance will likely be worse!

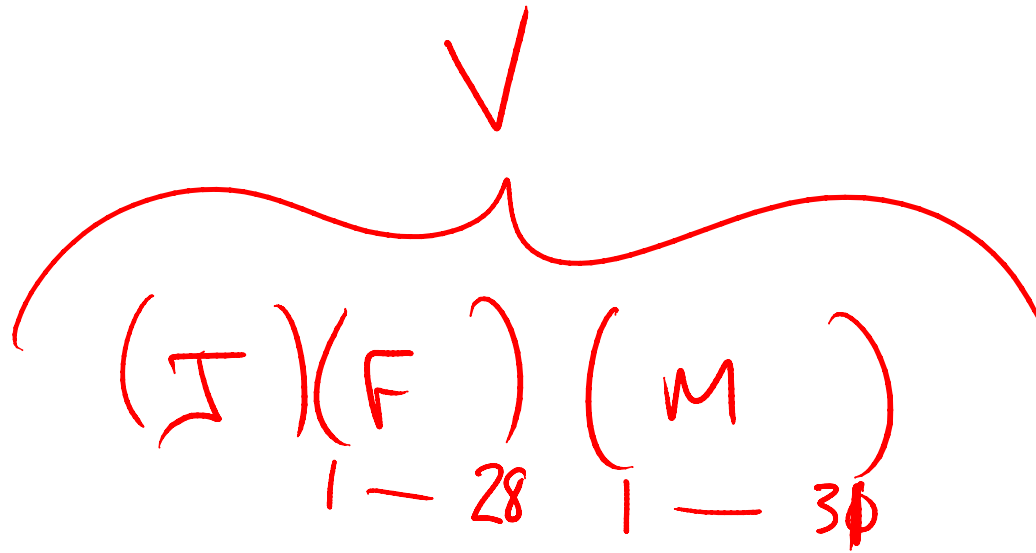
Check out the whitepapers on SODA (service-oriented database architectures).



M12

Partitioning in our module was all within a single database. However, SQL Server does support Scalable Shared Databases. These are RO databases that are attached to multiple servers and then balanced through WLBS.

[Scalable Shared Databases](#)

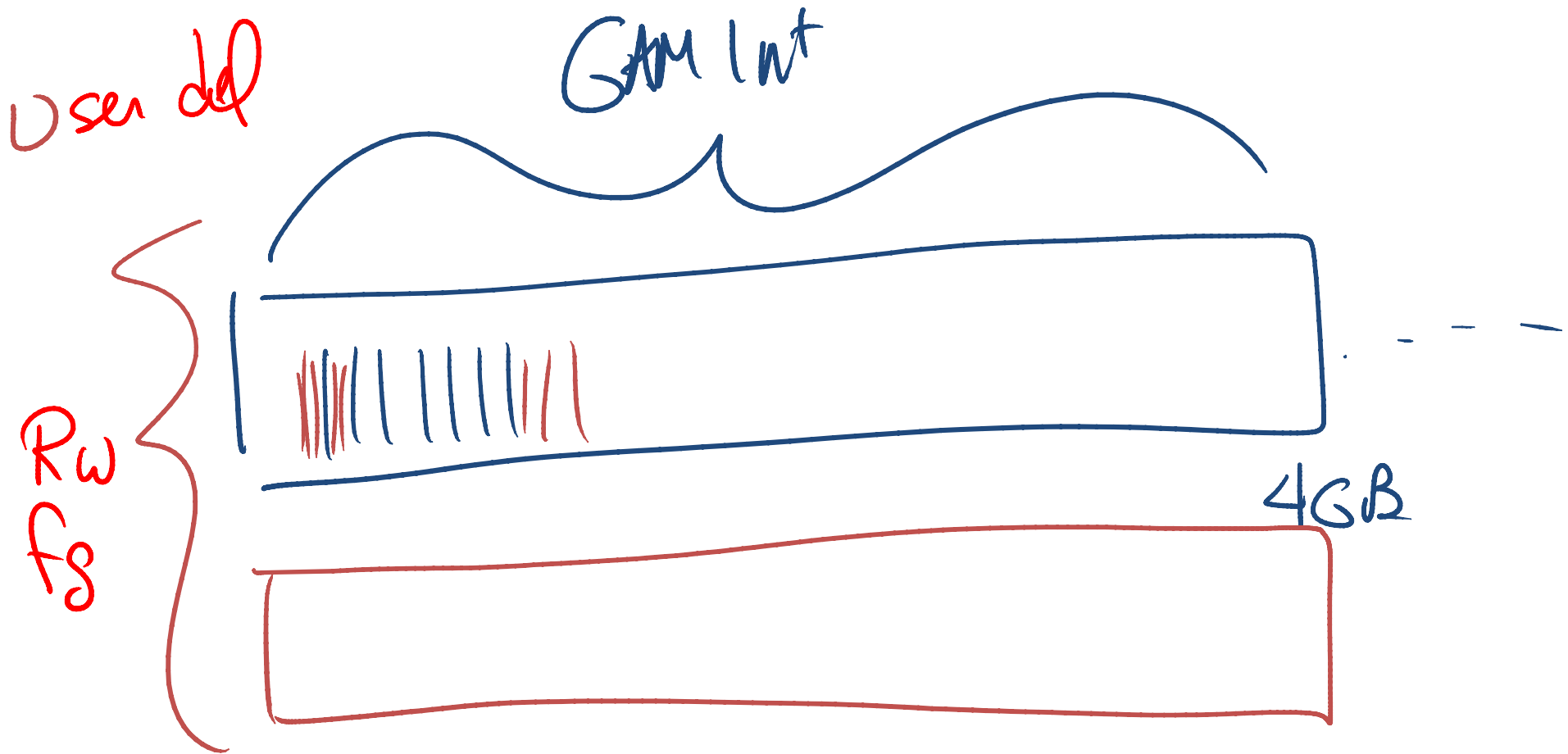


M12

Be careful with partitioned views... there's nothing that's going to test your business logic. As a result, if you miss a day (for example Feb 29, 2008) then inserts into the view will fail because there's no view that can store that date.

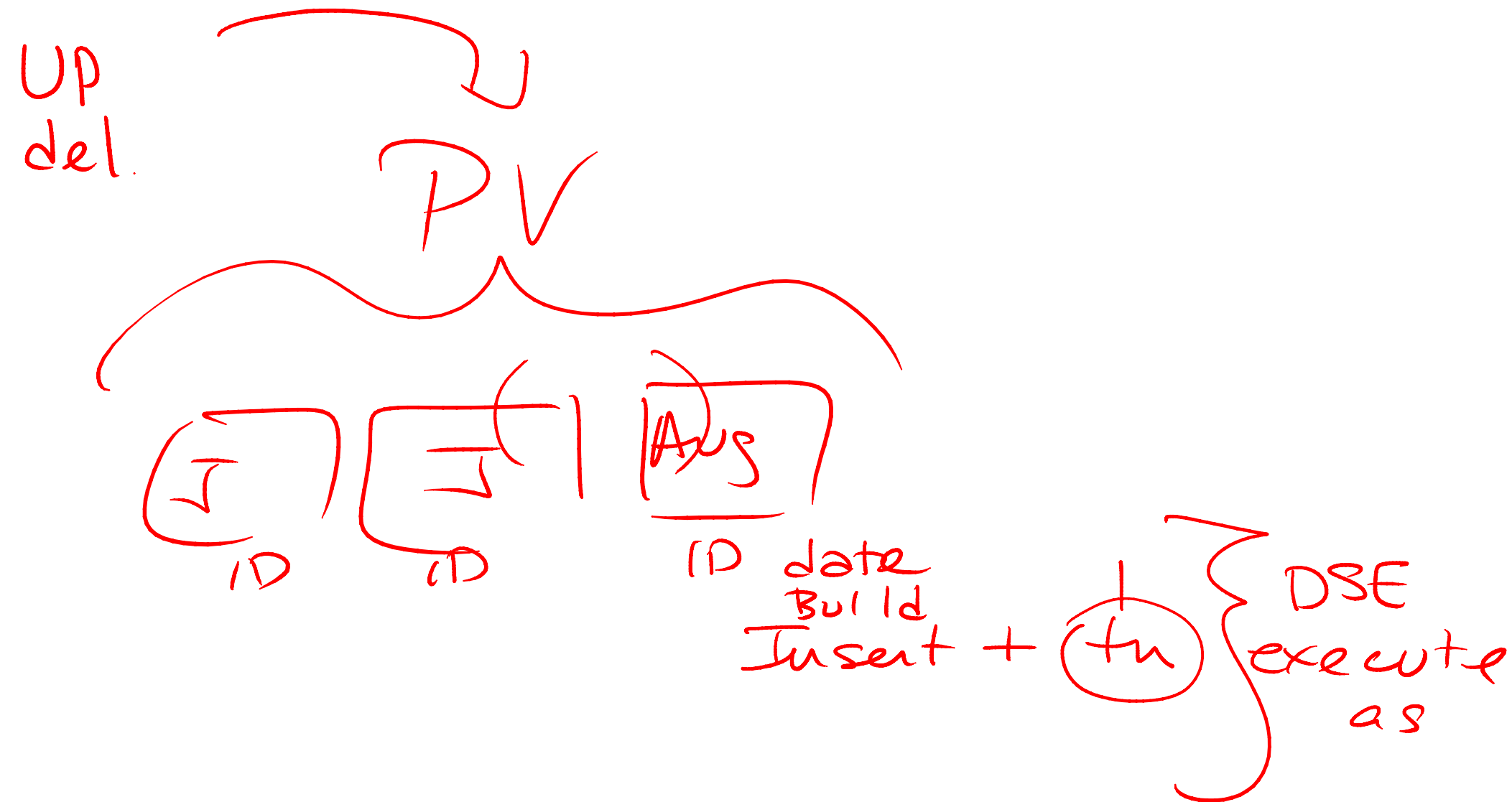
Alternatively, using partitioned tables – there's no way to have a gap or overlapping ranges.

But, there's a lot more to PVs and PTs (this isn't enough to choose one over the other) so this is just a bit more info to add to the list!



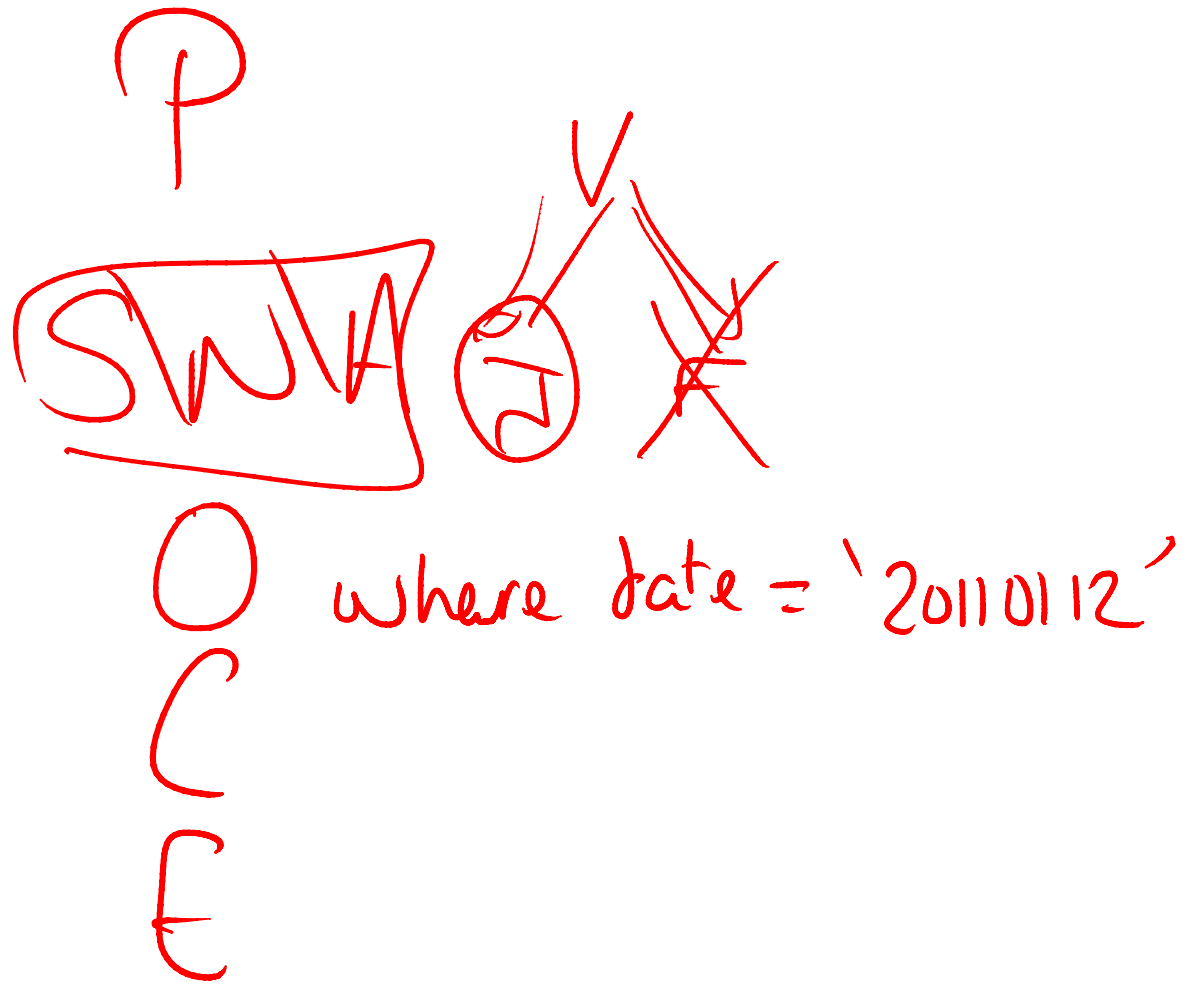
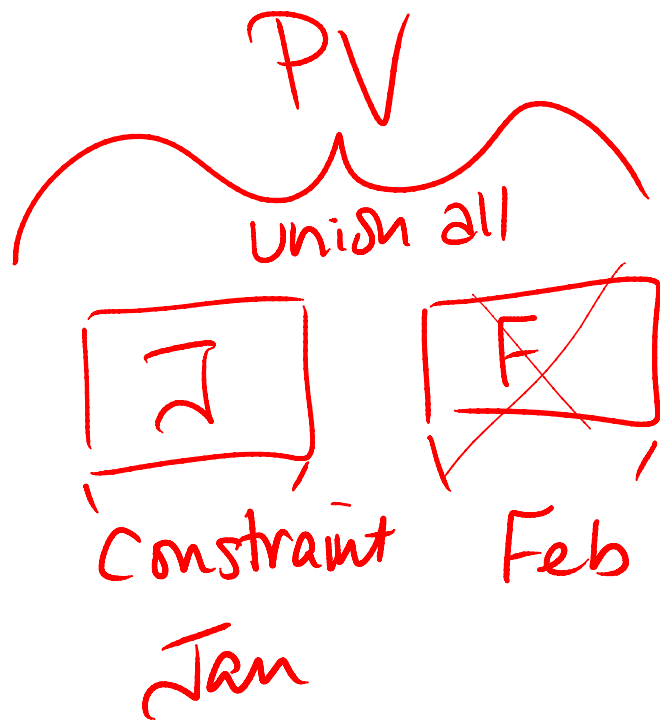
M12

If a RW filegroup has a lot of allocations then you can have contention (much like that in tempdb). One of the best strategies is to make sure that all RW filegroups have at least 2 files and as many as four (as a default practice). You might need more but just moving from 1 is already going to eliminate many problems.



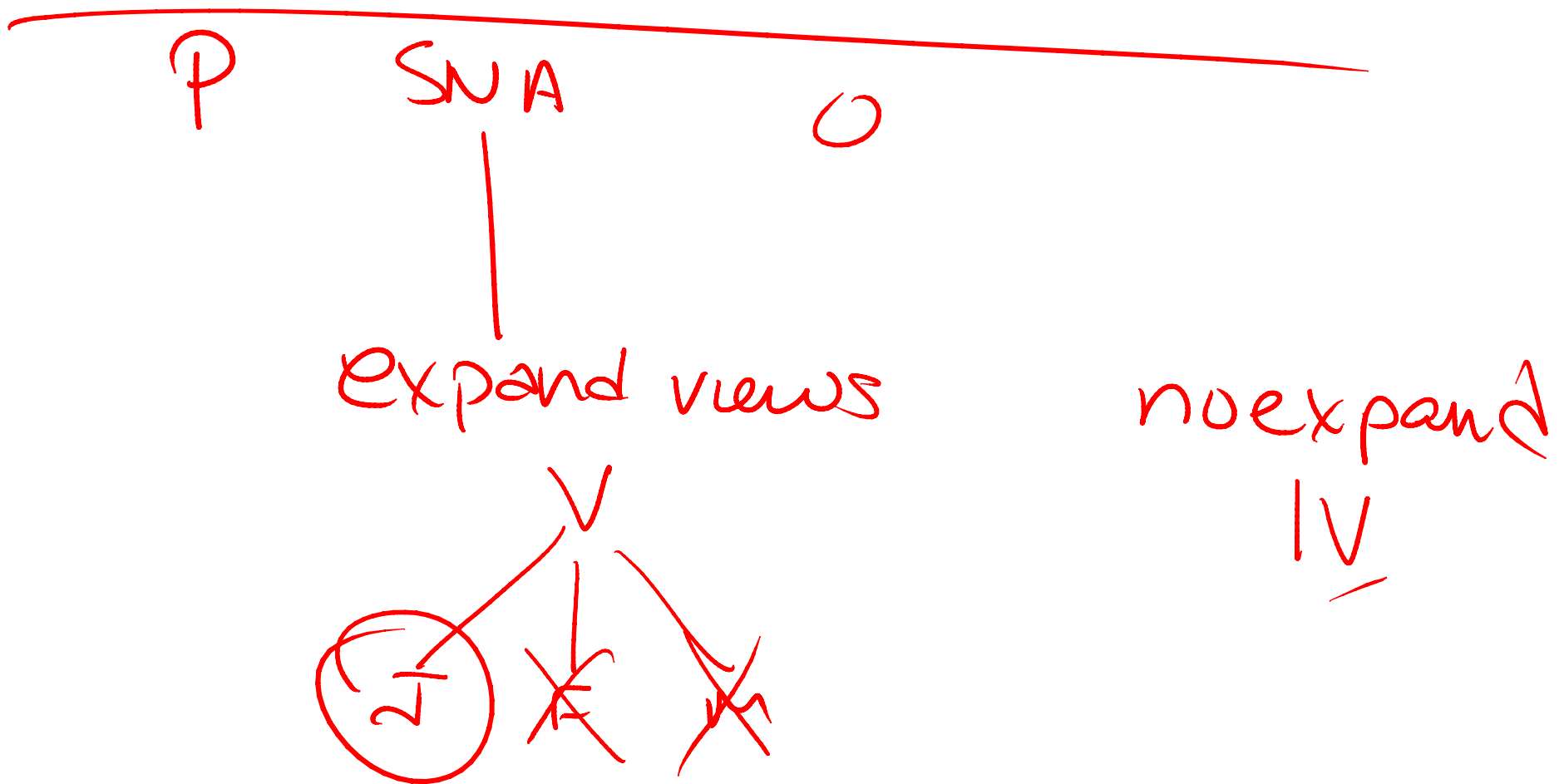
S15

Even with updateable partitioned views, I usually use application directed inserts. If you're concerned about dynamic string execution check out the Little Bobby Tables blog post: <http://www.sqlskills.com/BLOGS/KIMBERLY/post/Little-Bobby-Tables-SQL-Injection-and-EXECUTE-AS.aspx>



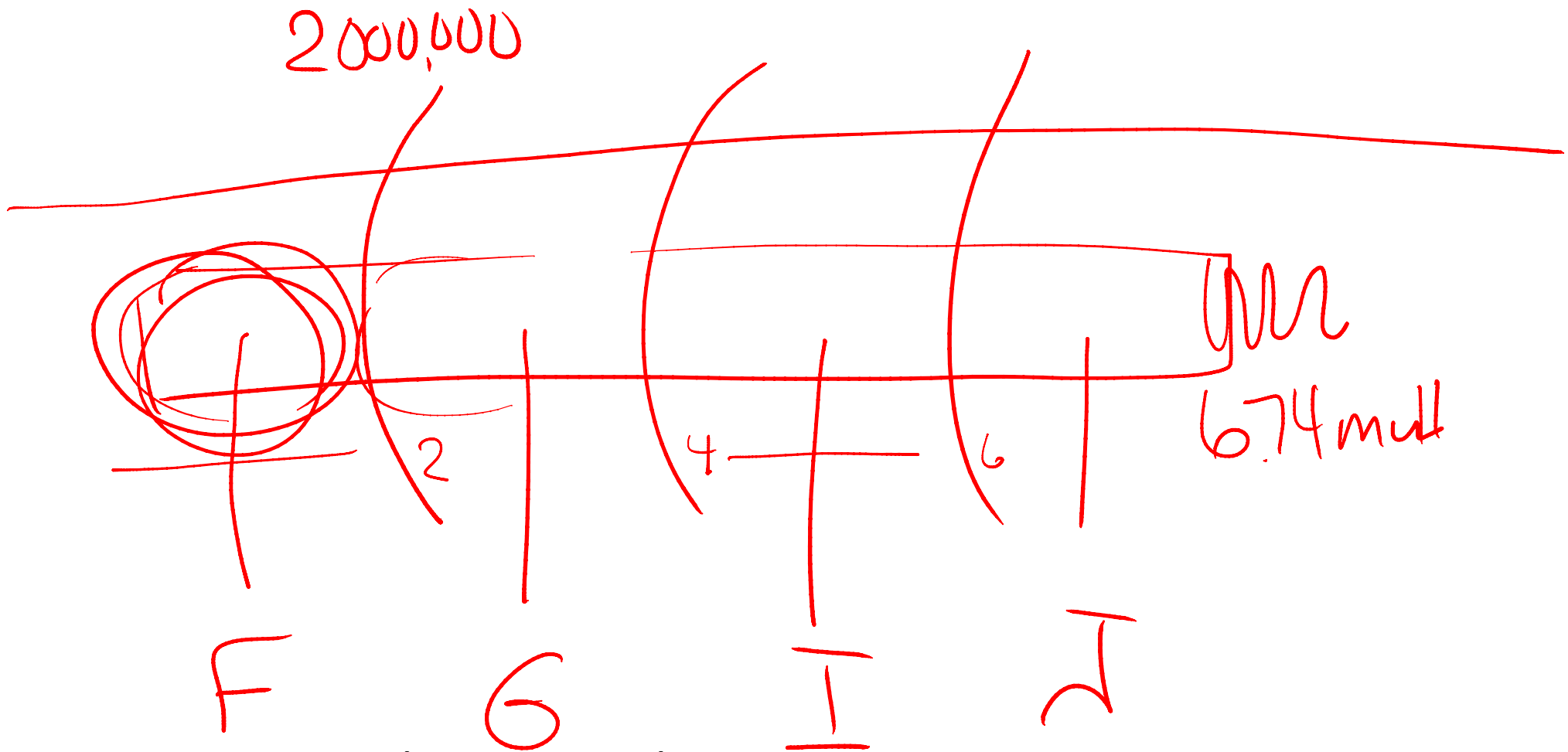
S17

Constraints are validated during optimization. SQL Server is able – when querying through a view – to generate the query tree and then “prune” the tree. This partition elimination removes any of the redundant tables from access. All of this is as long as the constraint is trusted (or, should I say as long as it’s NOT untrusted. 😊)



S17

Another way to look at this is that there are multiple phases of query processing. The second phase (standardization, normalization, algebraization) is where views are expanded to their base tables (known as EXPANDVIEWS). This might sound familiar because of the hint NOEXPAND – which is used to force SQL Server to use indexes on views (if for some reason SQL Server isn't using the IV).



Partitioning on integers (rather than date)

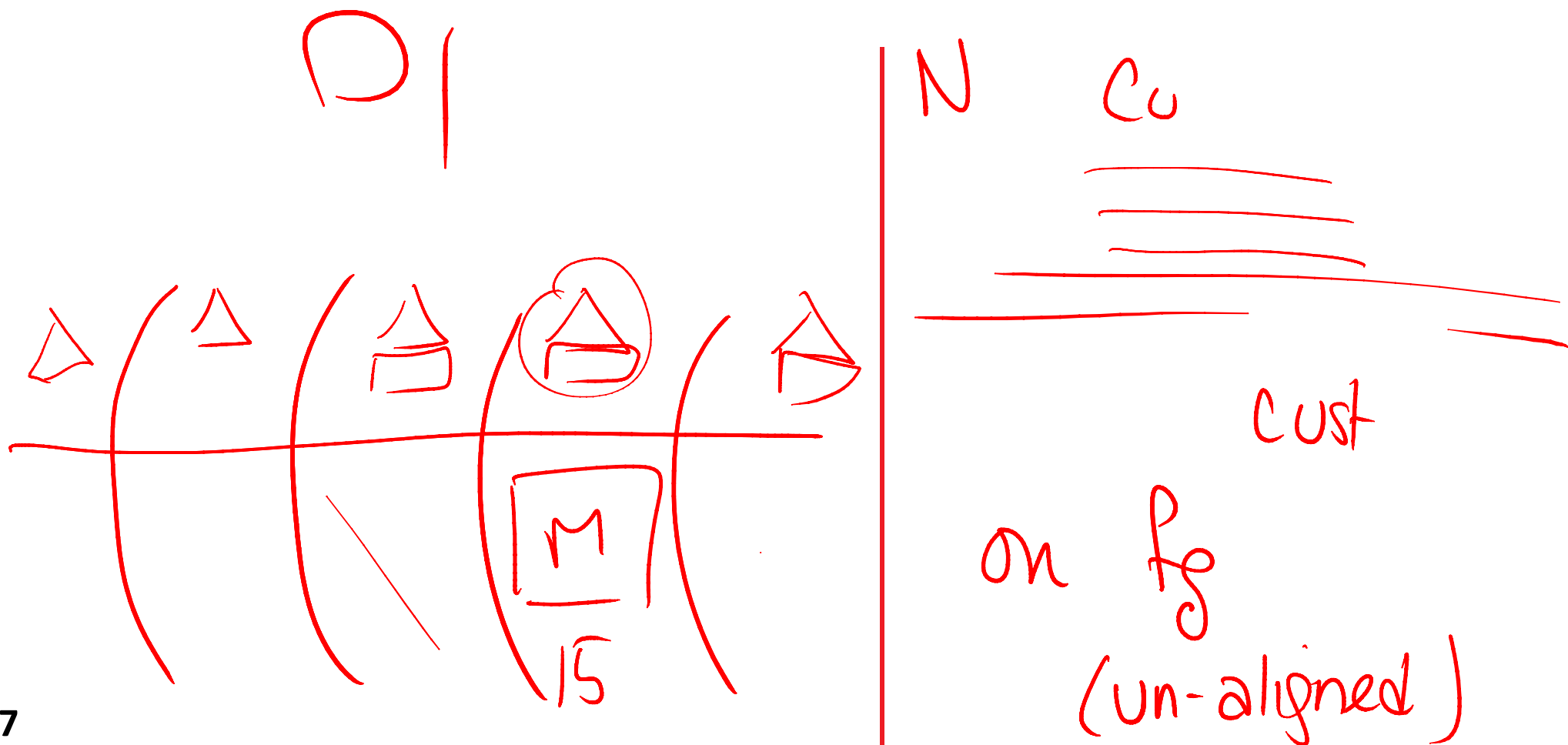
Another example of a partitioned table with 3 boundary points: 2million, 4 million and 6 million. It's a RIGHT-based partition function so the rows that match the boundary point will be to the RIGHT side. Specifically this translates into:

In partition 1: all rows less than 2 million

In partition 2: all rows equal to or greater than 2million and less than 4 million

In partition 3: all rows equal to or greater than 4million and less than 6 million

In partition 4: all rows equal to or greater than 6million



S37

For fast switching to be allowed – you must have ONLY aligned indexes. For an index to be aligned – there are rules. If the index is going to be unique (and aligned) then it MUST include the partitioning key. Indexes will default to being created on the same scheme as the table (they will default to being aligned – and therefore have all of these requirements).

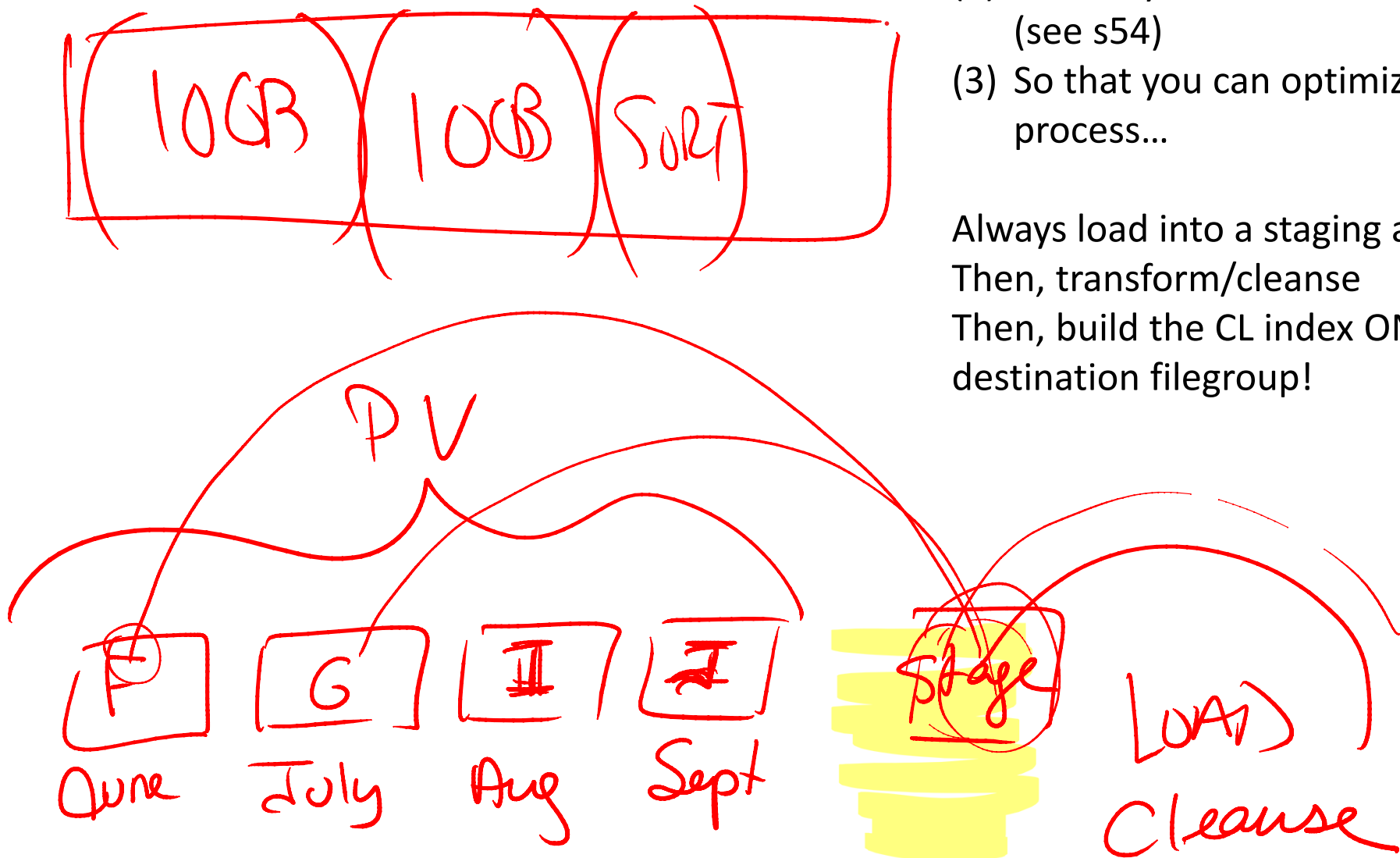
However, if you're not interested in fast switching – then you can have un-aligned indexes. In this case you create these indexes on a specific filegroup (or on a different scheme). These un-aligned indexes can be unique and do NOT have to include the partitioning column.

S44

Why do you need a staging area?

- (1) So that you don't need to overallocate space within destination filegroups
- (2) So that you don't have to shrink (see s54)
- (3) So that you can optimize the process...

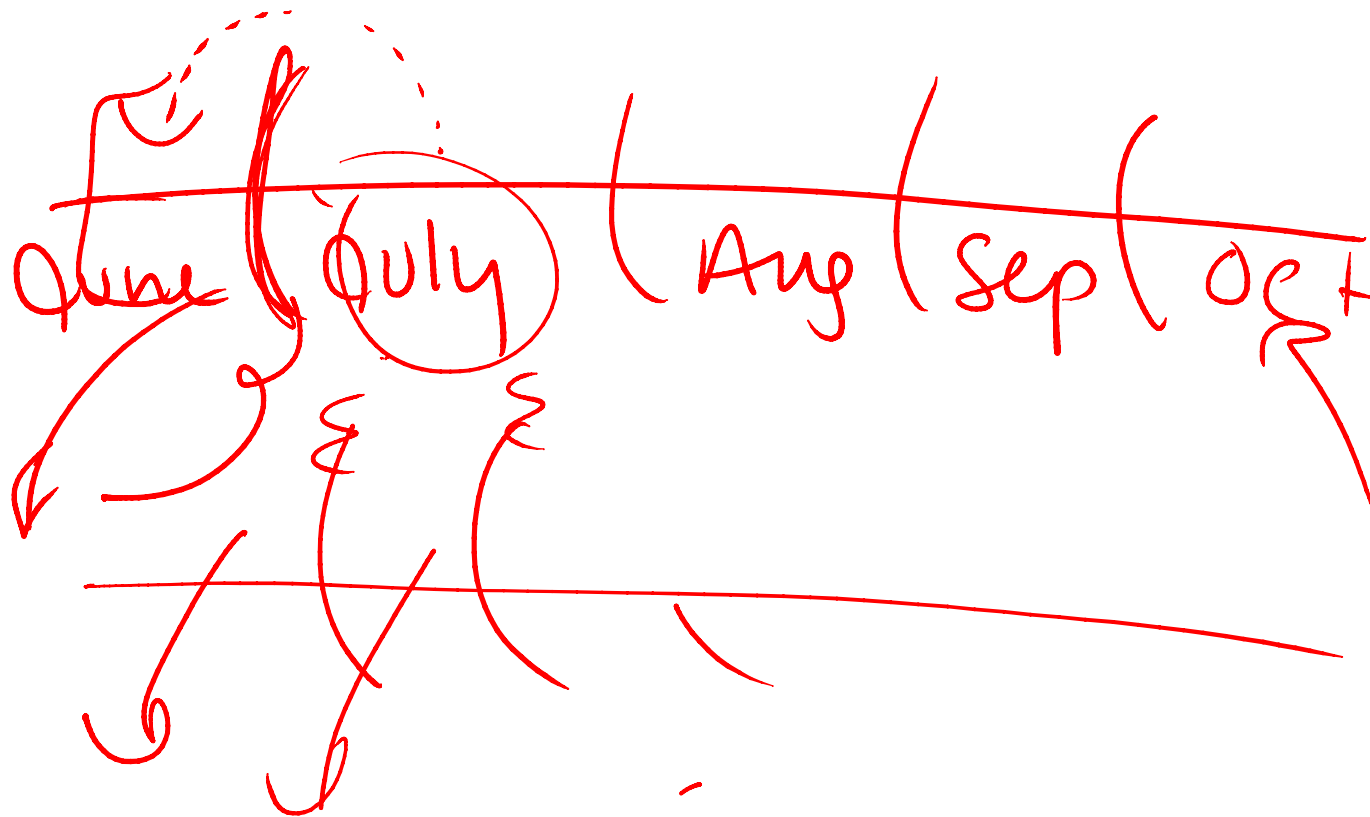
Always load into a staging area first.
Then, transform/cleanse
Then, build the CL index ON the destination filegroup!





S46

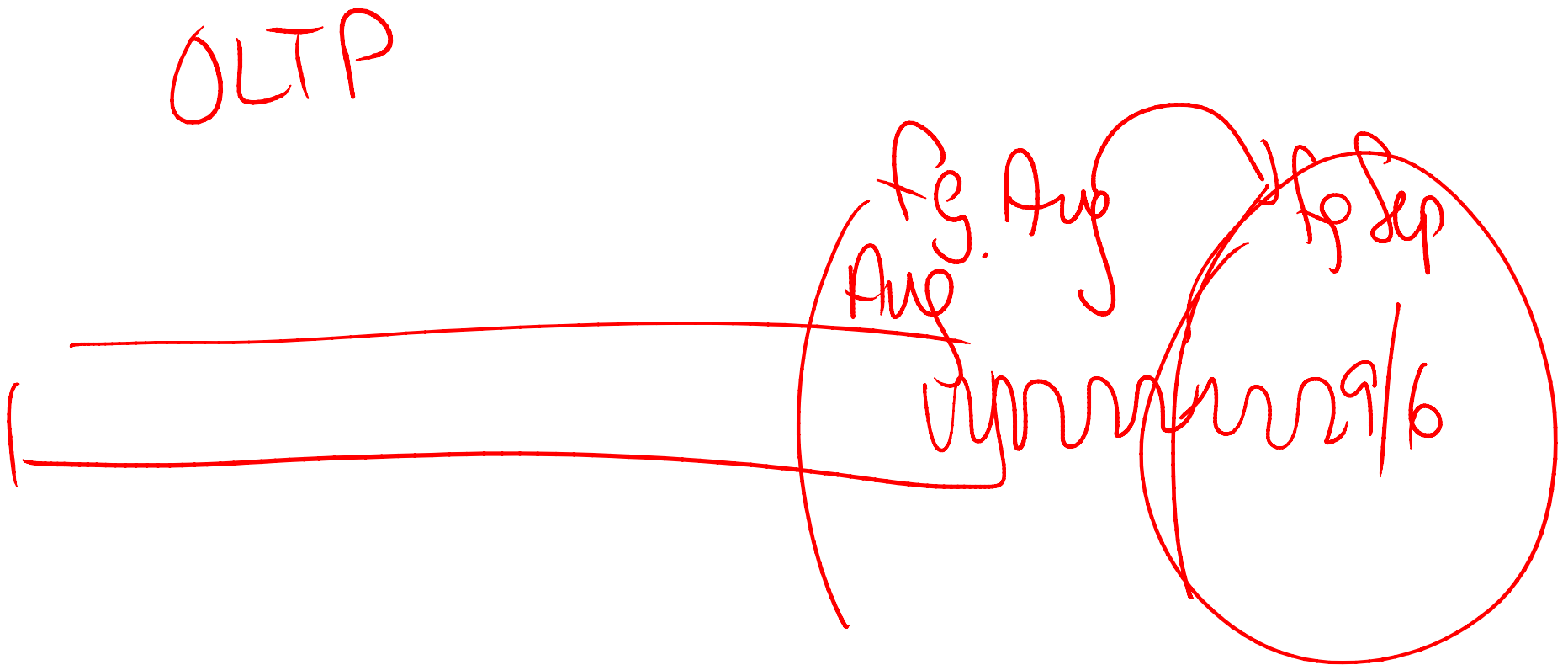
If the MERGE process is slow – you may have merged a boundary point that was NOT empty.



The most common case of this is when you start with a non-empty first partition using RIGHT-based partitioning. With RIGHT-based partitioning you should not MERGE until you have TWO empty partitions that surround the emptied boundary. Then, you can MERGE (top diagram).

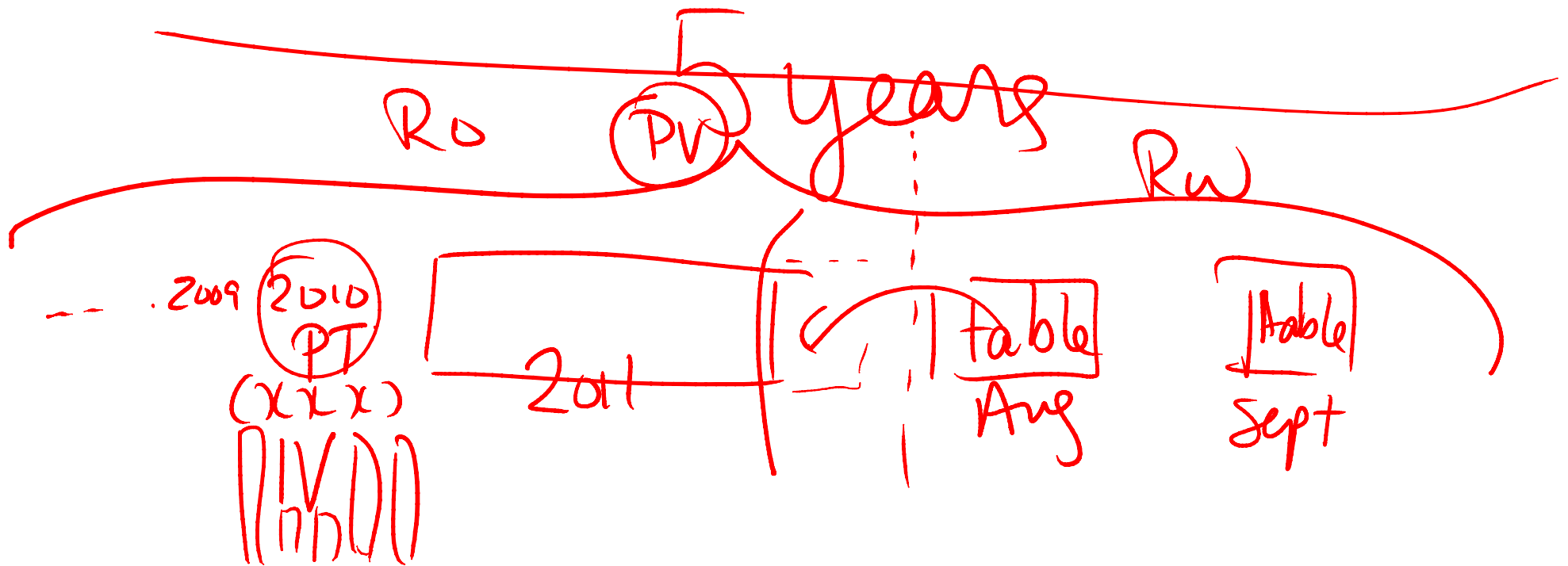
S48

With PTs – it's likely that your last partition (often, the current data) will be active. This is also where it's most likely that you'll have fragmentation. If you want to rebuild ONLY that last partition – you'll need take it offline to do it.



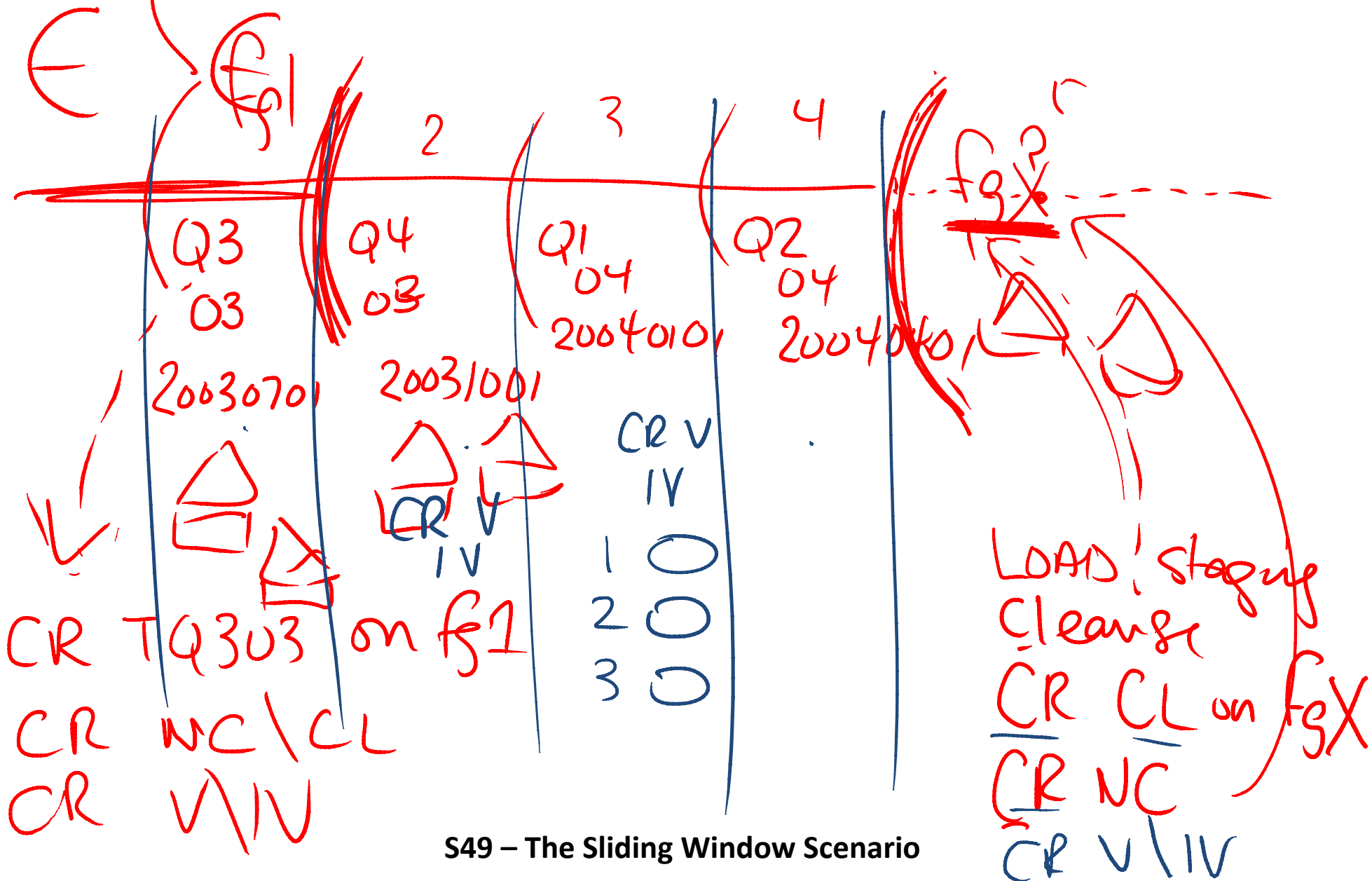
S48 (continued)

A better option is to separate the RW from the RO and put the RW in separate tables. Then, you can do online rebuilds at the table level. How do you deal with queries – put a partitioned view over them!



Scheme FGX
fn split '20040701'

merge



S49 – The Sliding Window Scenario

See the next slide for full details

The prior slide showed the sliding window scenario – but we went through it slowly:

(1) Preparing the table into which we'll switch OUT our old partition

Review all of the slides for the requirements here but the key point is that this “staging” table MUST be on the same filegroup as the partition you are going to switch out.

(2) Preparing the table for our data load and what will become our new partition to switch IN

Again, be sure to review all of the slides for the requirements here but one thing you need to make sure of is that there's a TRUSTED constraint on this table before you switch in.

(3) Change the partitioned table to support the new filegroup and data range

Always set the NEXT USED filegroup with ALTER SCHEME

Once you have the correct filegroup specified then ALTER FUNCTION...SPLIT

- Now, you're ready

(4) and (5) can be in any order. SWITCH IN the new partition and SWITCH OUT the old.

(6) Clean up

Merge the boundary point but ONLY if it's empty (the next picture will remind you of what happens when you don't MERGE an empty boundary point)

Backup the filegroup where the partition resides that you just switched out. OR, drop the table.

⑥ merge

③ Scheme next used fsg
split @ 20040701

E E

fsg1

fsg2

fsg3

fsg4

fsg5

fsg6

Switch
OUT

July
03

Oct
03

Jan
04

Apr
04

Switch
IN

⑤

④ CR T 200307Q1

≡
≡
≡
≡

on fsg2

①

CR CL

CR NC

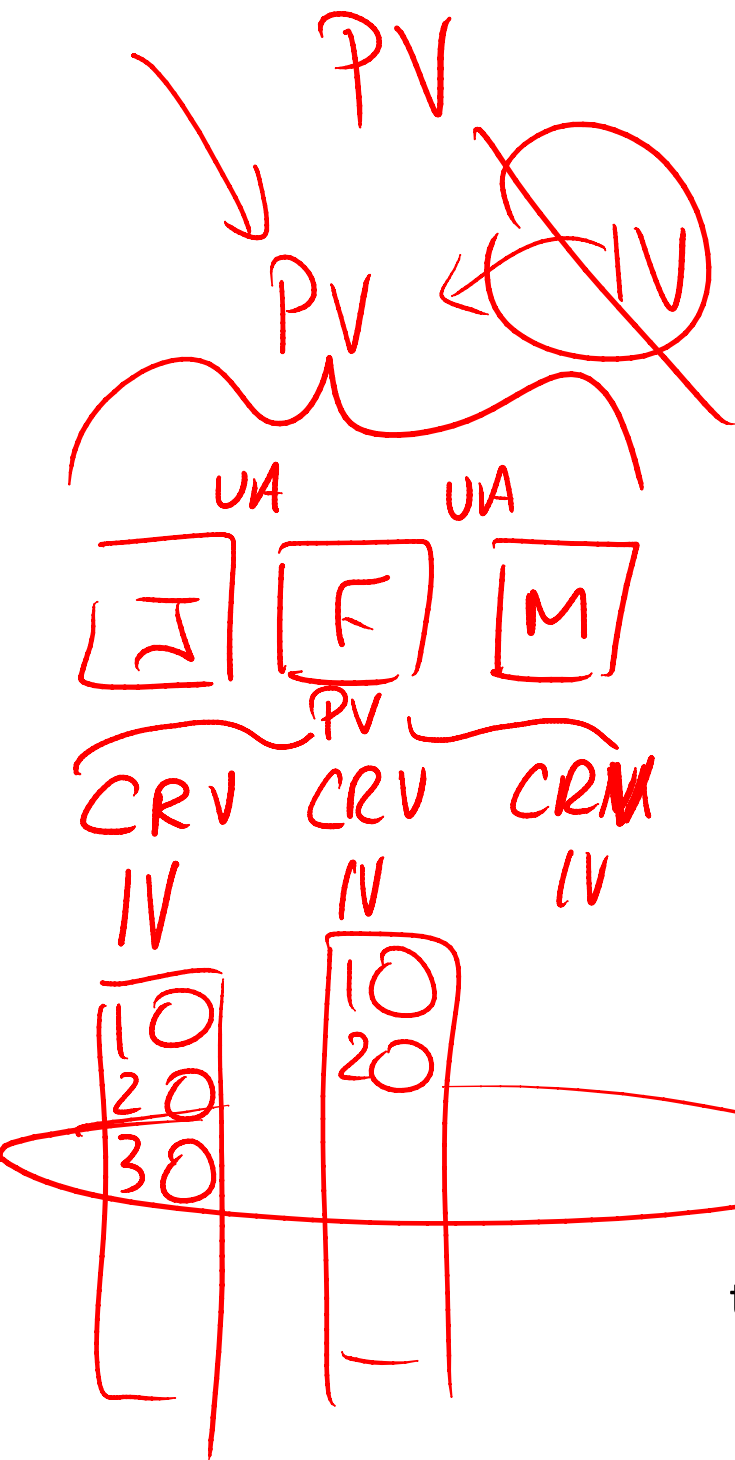
2008 → CR V/IV

July 04

Const.

② Staging
LOAD

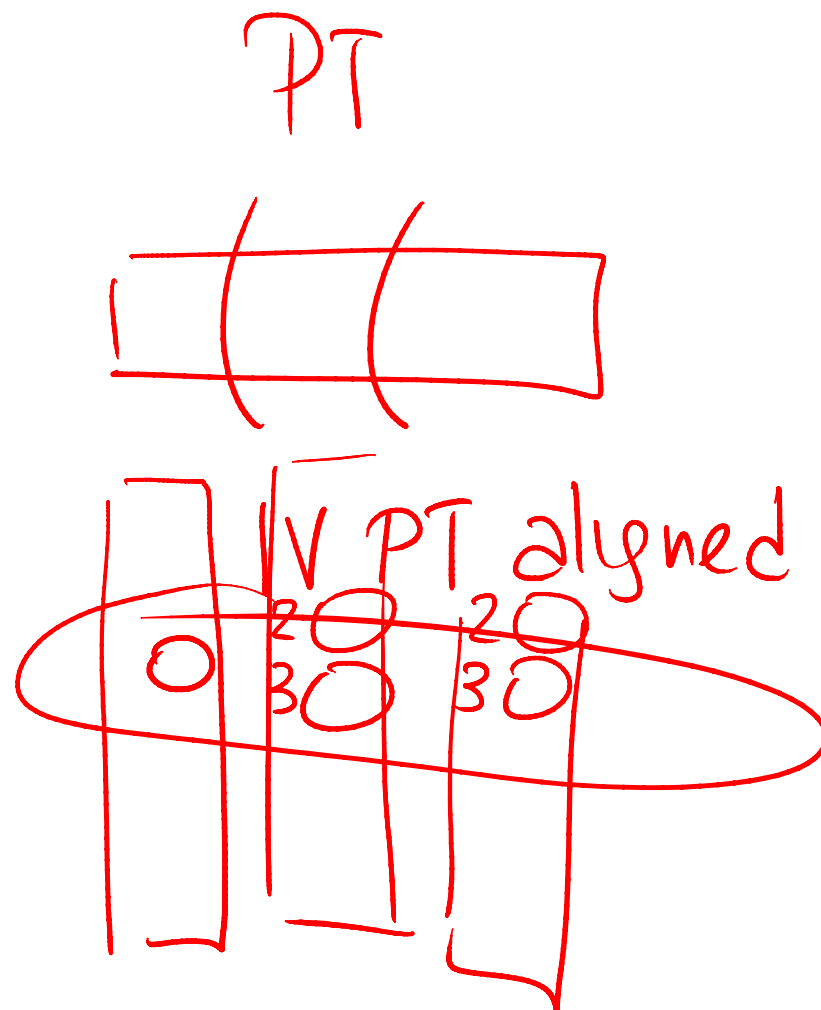
cleanse
CR CL on fsg6
CR NC
CR V/IV

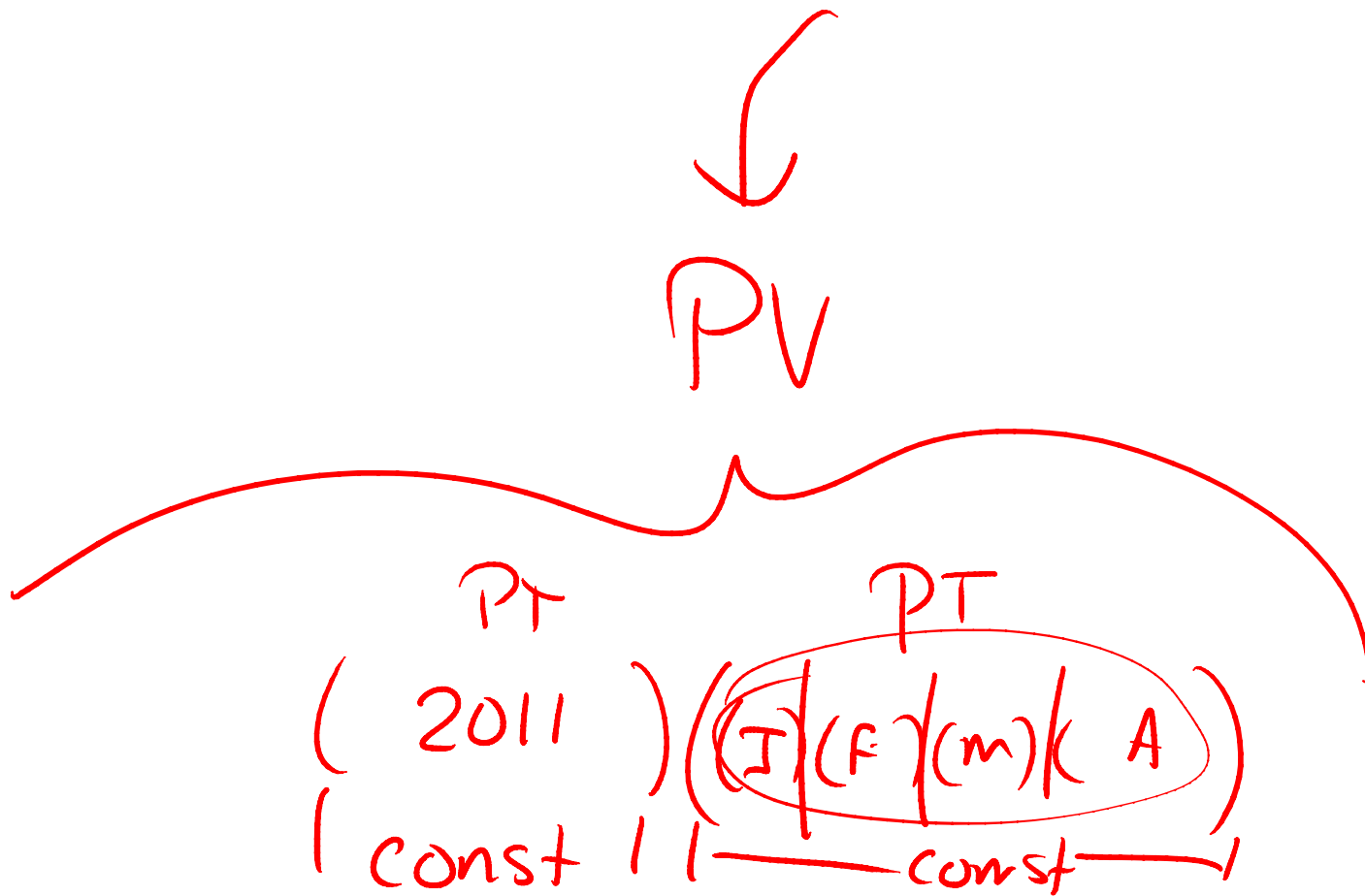


With PVs - If you want to create indexed views, you must create them individually per table (for the tables that underpin the PV). If you're not on EE then you'll also need to create a specific view to access them with the (WITH NOEXPAND) hint.

S50

With PTs - If you want to create indexed views, you must create them as partition-aligned (for fast switching). This is a new feature of SQL Server 2008+





M12

You can have partitioned tables and/or non-partitioned tables that underpin a view. The only requirement – make sure that each of the tables (UNIONed in the view) has a constraint!x

SQLskills Immersion Event

IE1: Internals and Performance

Module 3: Locking and Blocking

Kimberly L. Tripp
Kimberly@SQLskills.com



Implicit / auto commit trans.

Set implicit_transactions on

Update

Insert

Commit tran

Explicit
Begin Tran

Commit Tran

S3

While discussing locking basics, we talked about ACID properties.

We also started to discuss transactions:

- *Implicit* is now auto-commit. In most people's minds implicit used to mean that SQL Server implicitly treats your statements as a transaction. However, Oracle's implicit begin a transaction for you and so these should not be confused. To reduce confusion, most people now use the term auto-commit instead.
- Explicit are user-defined transactions where you've explicitly defined the begin/commit.

ACID

Isolation

explicit [Begin tran
Commit tran

implicit $\Rightarrow$ Oracle

auto commit Set implicit-transactions on

S3

While discussing locking basics, we talked about ACID properties.

This picture is described better on the following new picture/slide.

Transaction Properties

ACID

→ **Durability** – covered by write-ahead logging.

→ **Isolation** – this is what you can control the most. What do readers see?

→ **Consistency** – everything affected by the transaction (triggers, indexes, constraints) will remain consistent as transactions take data from one state to the next.

→ **Atomicity** – A transaction is a single, logical, unit of work. “All or nothing.”

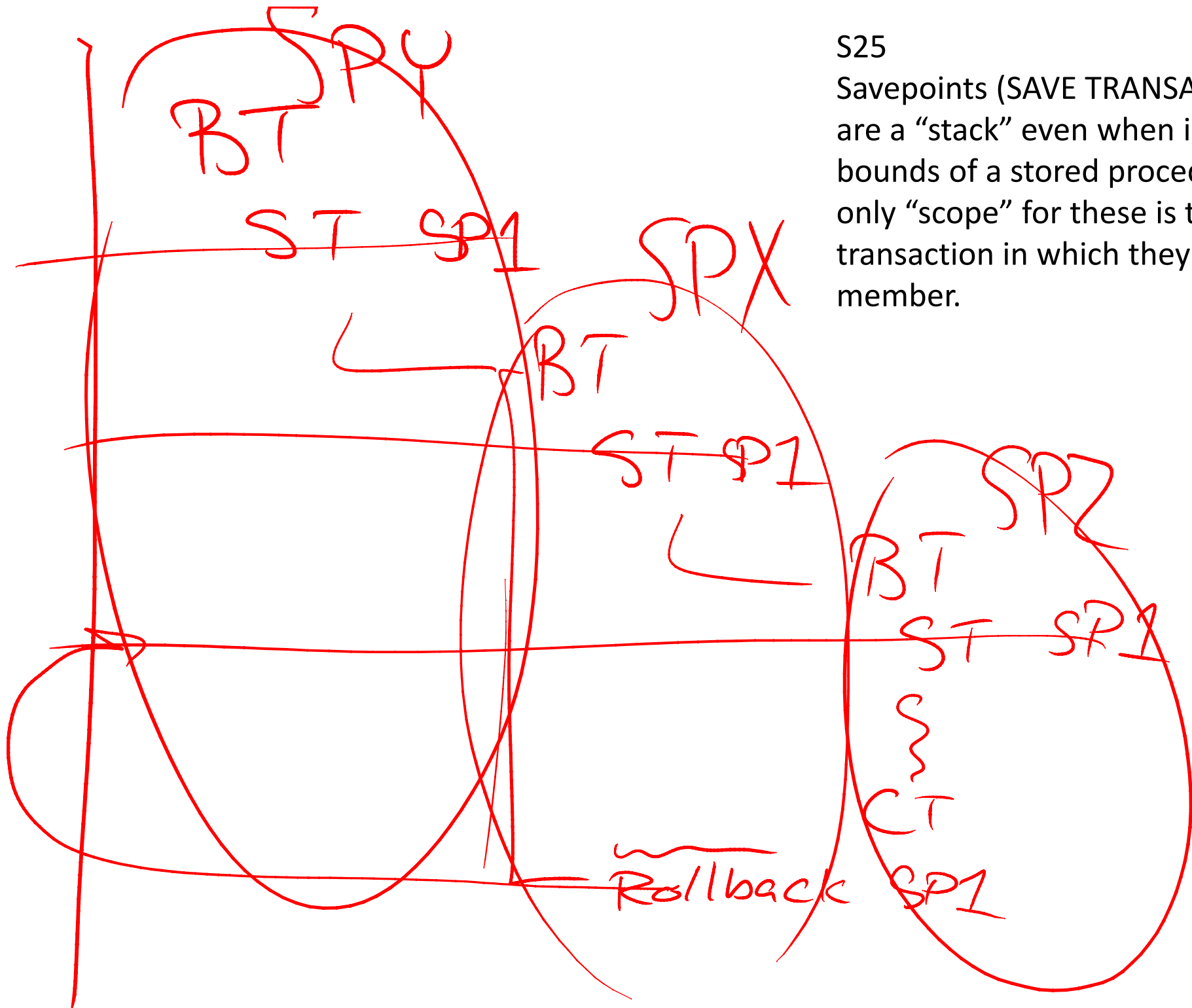
- **Implicit (or auto-commit) transactions** – each statement is a transaction. If a single statement affects multiple rows then ALL rows shall be affected or NONE of those rows will be affected. SQL Server treats EVERY statement as an auto-commit (or implicit) transaction without requiring a BEGIN or COMMIT. If the same behavior as Oracle is desired: SET IMPLICIT_TRANSACTIONS ON. Any statement will begin the transaction; COMMIT TRANSACTION is required to complete.
- **Explicit transactions** – can widen the boundaries of your transaction by explicitly stating a BEGIN TRANSACTION and a COMMIT TRANSACTION. Be careful though – some statements errors only cause the statement to be rolled back and not the transaction (unless SET XACT_ABORT ON).

From BOL:

- When SET XACT_ABORT is ON, if a Transact-SQL statement raises a run-time error, the entire transaction is terminated and rolled back.
- When SET XACT_ABORT is OFF, in some cases only the Transact-SQL statement that raised the error is rolled back and the transaction continues processing. Depending upon the severity of the error, the entire transaction may be rolled back even when SET XACT_ABORT is OFF. OFF is the default setting.

S25

Savepoints (SAVE TRANSACTION) are a “stack” even when in the bounds of a stored procedure. The only “scope” for these is the “one” transaction in which they are a member.



S27 – aka Transaction Madness

The key points are:

BEGIN TRANSACTION

- Increments @@trancount

COMMIT TRANSACTION

- Decrements @@trancount

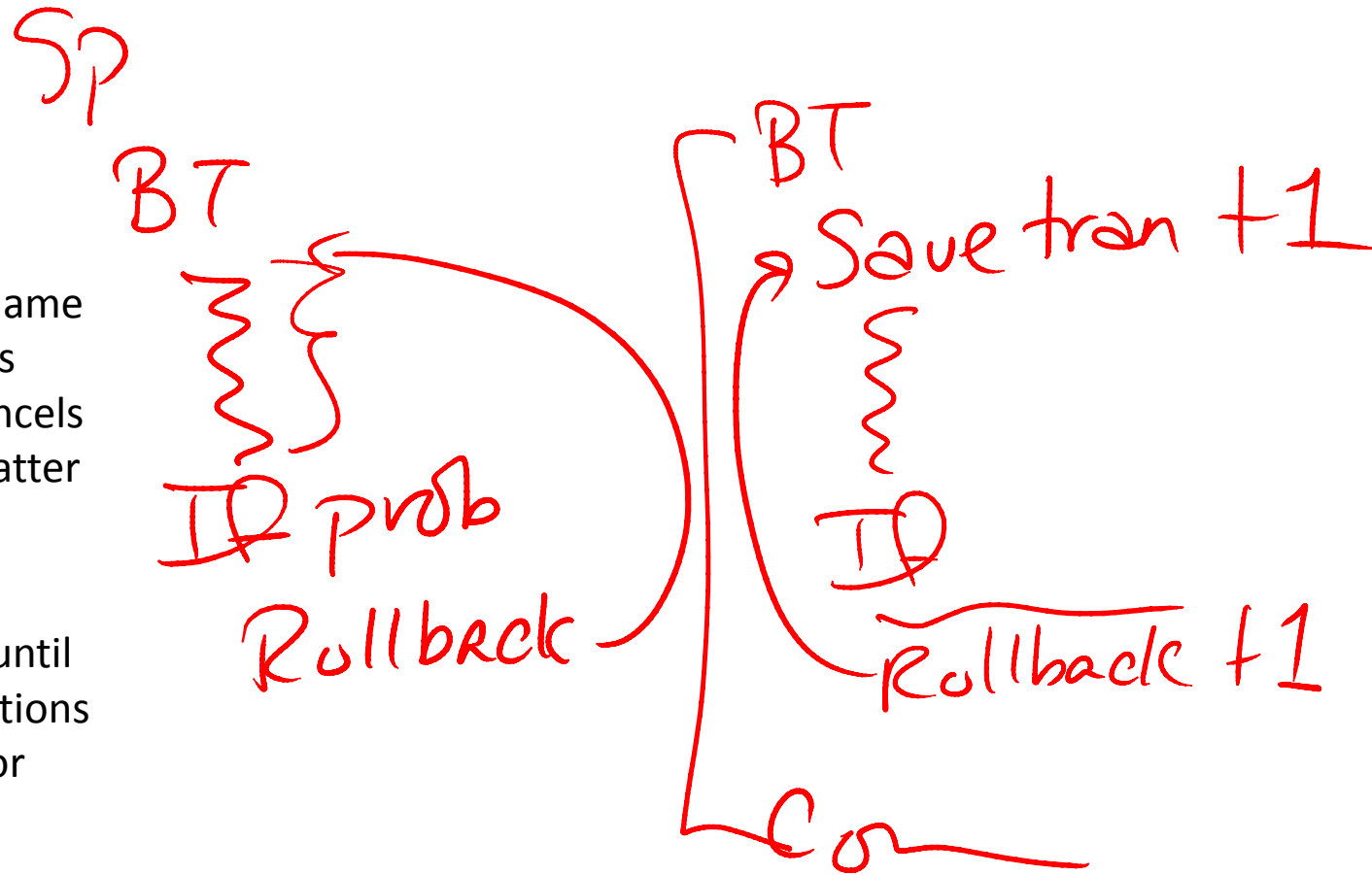
ROLLBACK TRANSACTION

- When used with a transaction name OR without a name at all – always resets @@trancount to 0 and cancels ALL pending transactions – no matter how deeply nested.

A transaction is NOT committed until @@trancount = 0. When transactions are nested you MUST COMMIT for every BEGIN

Savepoint (SAVE TRANSACTION NAME) do NOT affect @@trancount and can be rolled back to safely without affecting @@trancount.

A stored procedure is NOT a transaction in and of itself. You MUST use BEGIN/COMMIT to engage the ACID properties of the statements of a stored procedure.



Be sure to go through the demo scripts if you need more insight into transactions!

TRAN BT 1

Se _____

U _____

S25

Savepoints (SAVE TRANSACTION) are a "stack" even when in the bounds of a stored procedure. The only "scope" for these is the "one" transaction in which they are a member.

Save tran SP1

{

exec p1

BT 2

{

Save tran SP2

{

EXEC SP2

CT 1

stack

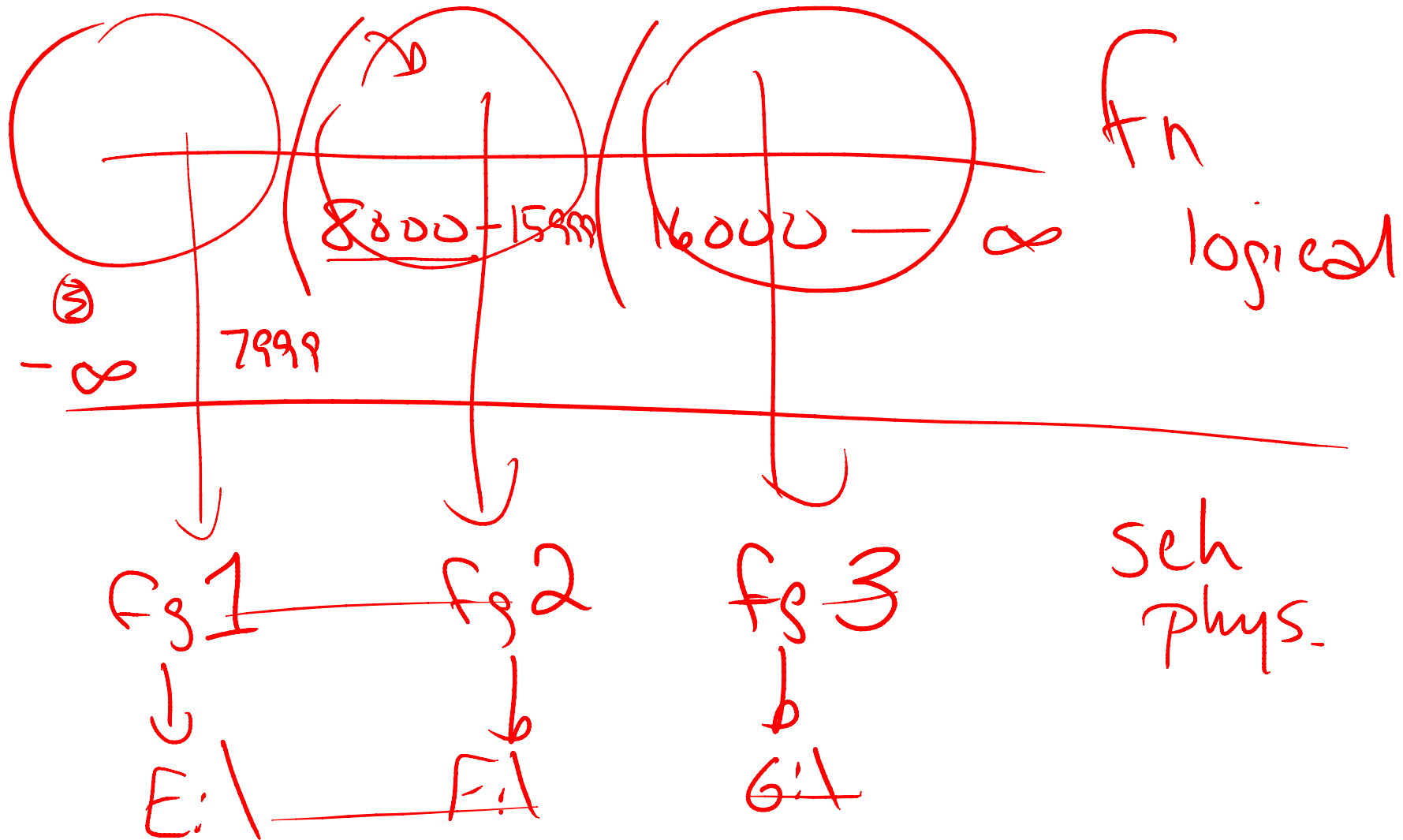
The rollback that's below actually rolls back to the LAST savepoint (SP1) in the stack. You'll want to make sure that every save point is unique and well-named.

IP

ROLLBACK TRAN SP1

{

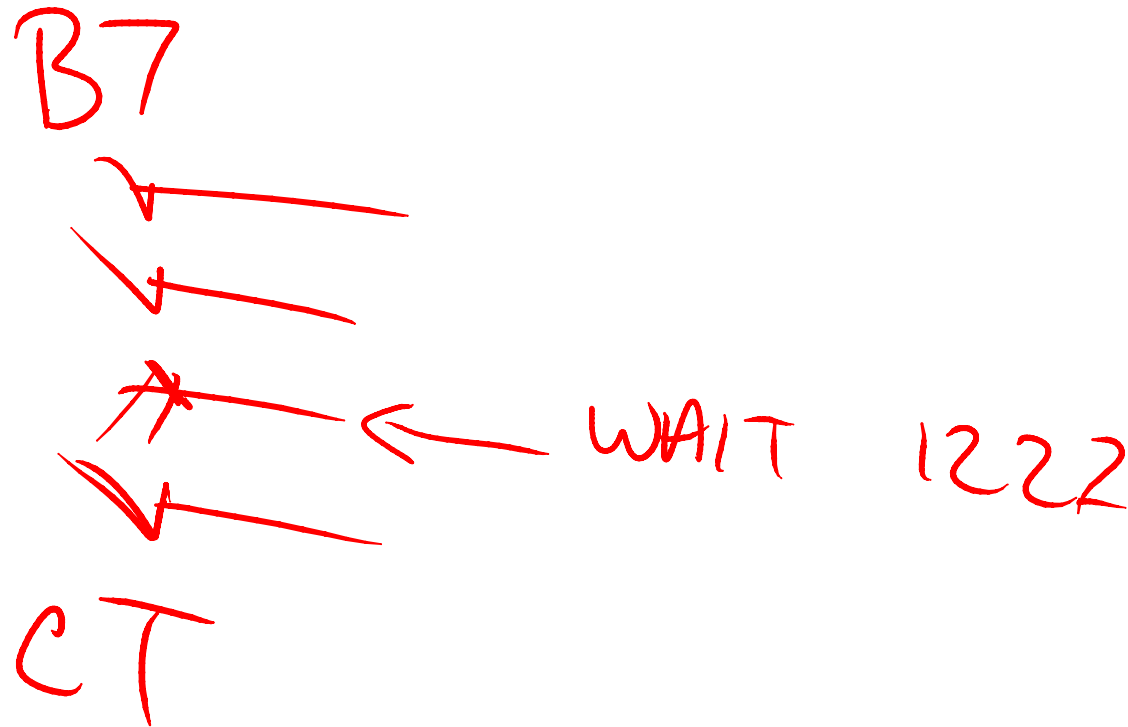
CT 2



S41

This is from the lock escalation demo. This picture shows how the boundaries define the logical placement of data within the 3 partitions. All this pic is describing (briefly) is that the two RIGHT boundary points of 8000 and 16000 define the breakdown as follows:

$-\infty$ to 7999	8000 to 15999	16000 to ∞
Partition 1	Partition 2	Partition 3



S46

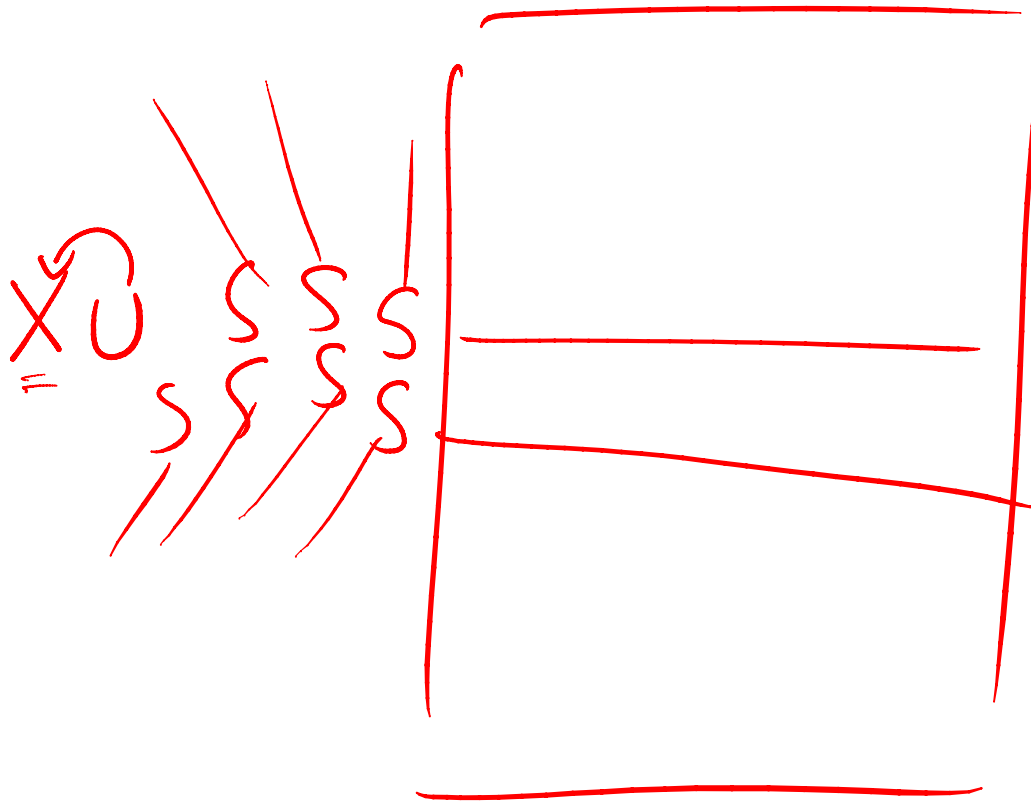
If a statement of a transaction hits a USER-DEFINED error – what happens?

It's up to the application to determine the fate of the transaction.

In this case, if the 3rd statement errors with Error 1222 (Lock request time out period exceeded.) then the statement will be rolled back but the transaction will be pending.

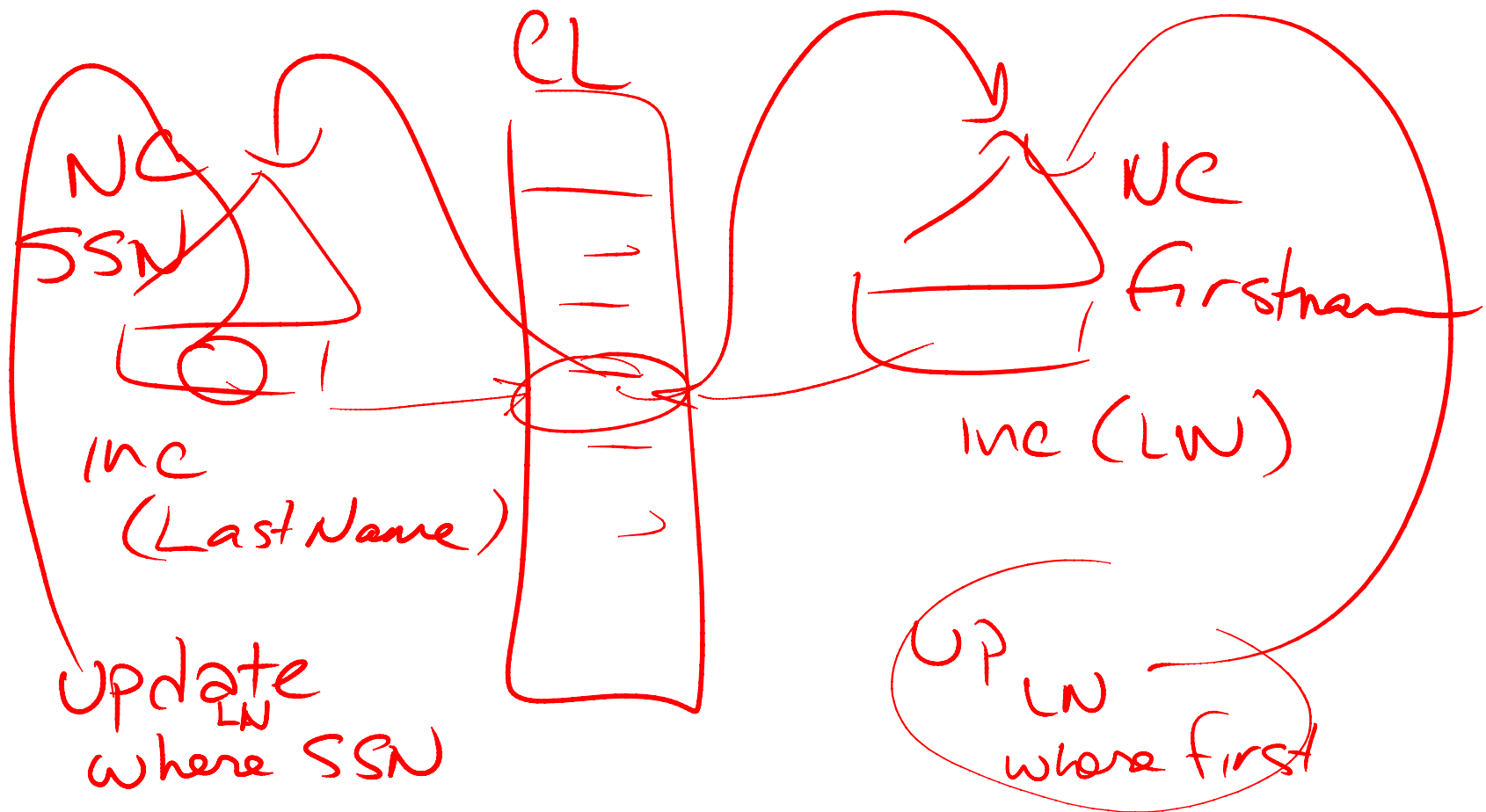
This is the default behavior when SET XACT_ABORT is not on.

Having better ~~error~~ error handling – writing code within TRY/CATCH blocks is essential for reducing errors and increasing data integrity.



S48

This was just a further discussion about the process that occurs when the update lock is converted to an exclusive lock. The discussion really focused back on the anatomy of a data modification. The conversion does not occur until SQL Server has to process the row. So, in a large/longer running transaction, you might have many update locks acquired and then as the data is processed, the rows are converted. Until the conversion is requested, shared locks are also allowed on these rows (even while the update lock is waiting) because of relaxed fifo. However, once the conversion is requested then ALL locks will wait.



S55

This picture described a deadlock that can occur when multiple indexes exist on a single table and they INCLUDE columns that are being updated.

CL = clustered table

Left side = index on SSN INCLUDE (lastname)

Right side = index on firstname INCLUDE (lastname)

Imagine an update to firstname that uses WHERE SSN = running at the same time as an update to lastname that uses WHERE Lastname = (let's not debate how bad this latter update is ☺).

These updates are MUCH more likely to create deadlocks because of how locking works within indexes. To reduce the potential here – I often use sp_tableoption.

BT

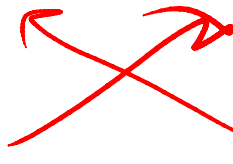
Upd saving

Upd check

BT

Up check

Up saving



Upd check
Upd saving

Upd check W
Upd saving

S60

Deadlocks often occur when resource patterns are interleaved. Where possible, rewriting code can be really helpful. A simple example is “the banking transaction” where one user is moving money from checking to savings and the other from savings to checking. This “crisscross” of activity will be more prone to deadlocking. One option (because this is a transaction) is to reorder the statements of the transactions to always access the tables in a particular order. If this were the case then these deadlocks would be removed and replaced with blocking (not deadlocks).

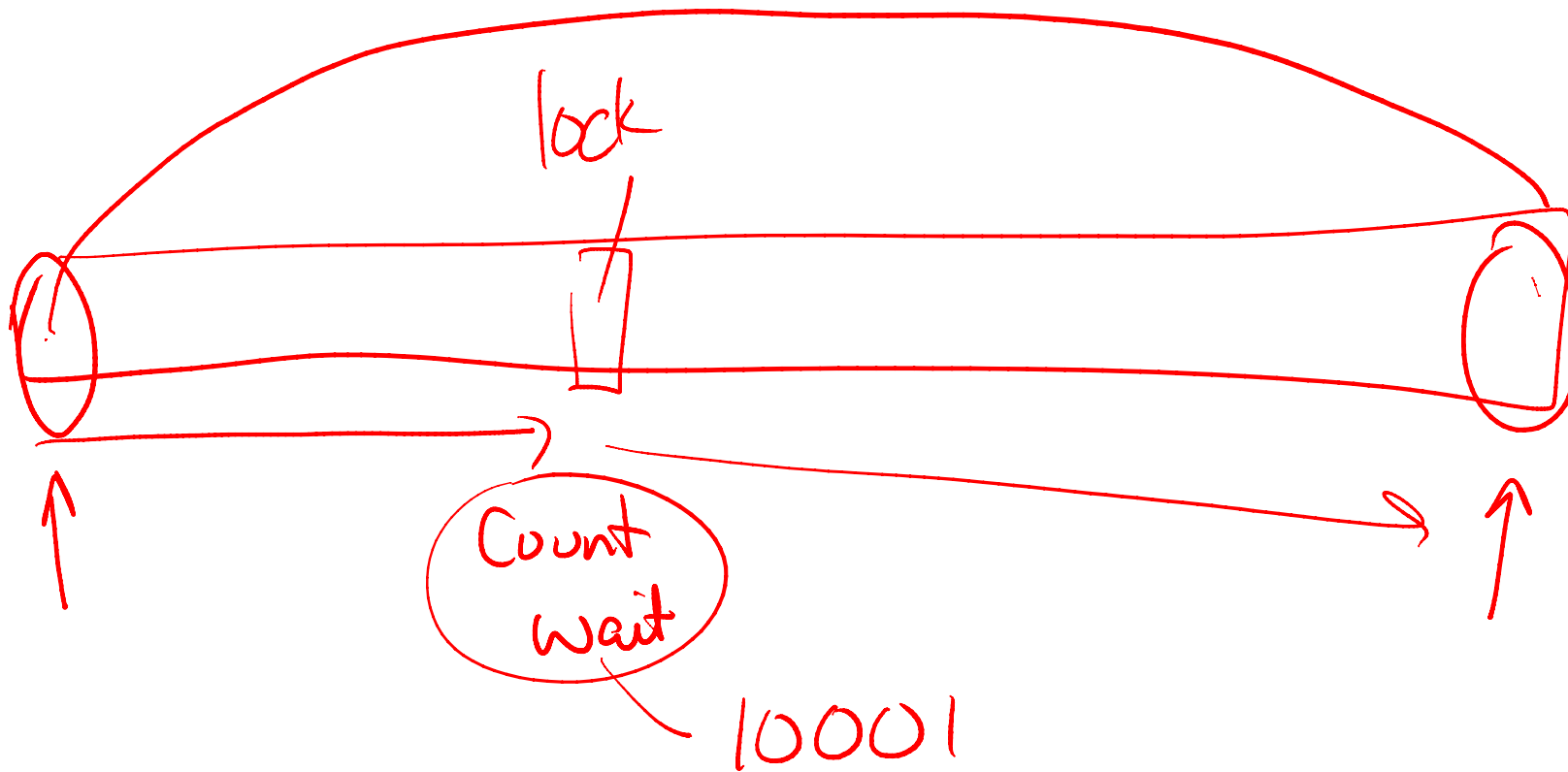
SQLskills Immersion Event

IE1: Internals and Performance

Module 4: Snapshot Isolation

Kimberly L. Tripp
Kimberly@SQLskills.com





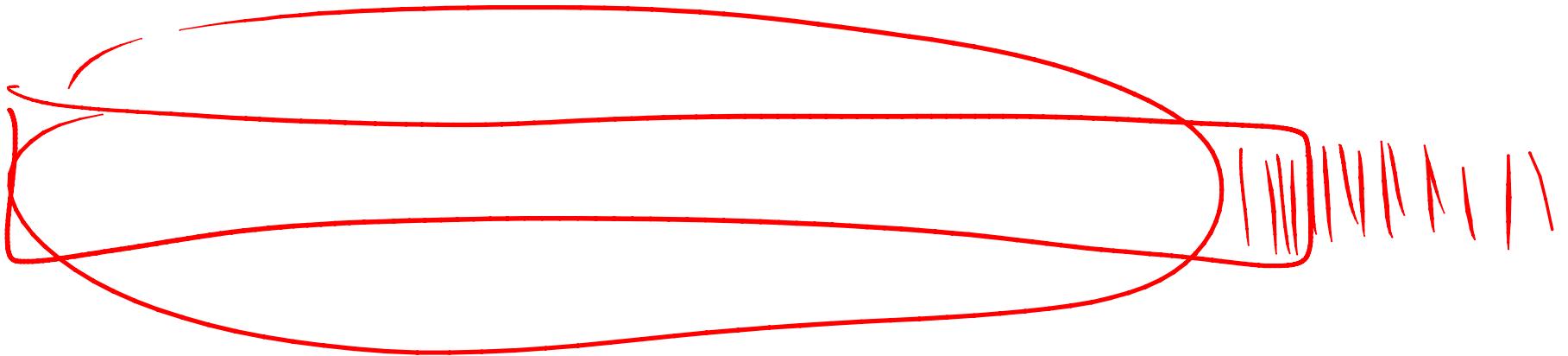
Demo: Non-repeatable reads (Inconsistent Analysis)

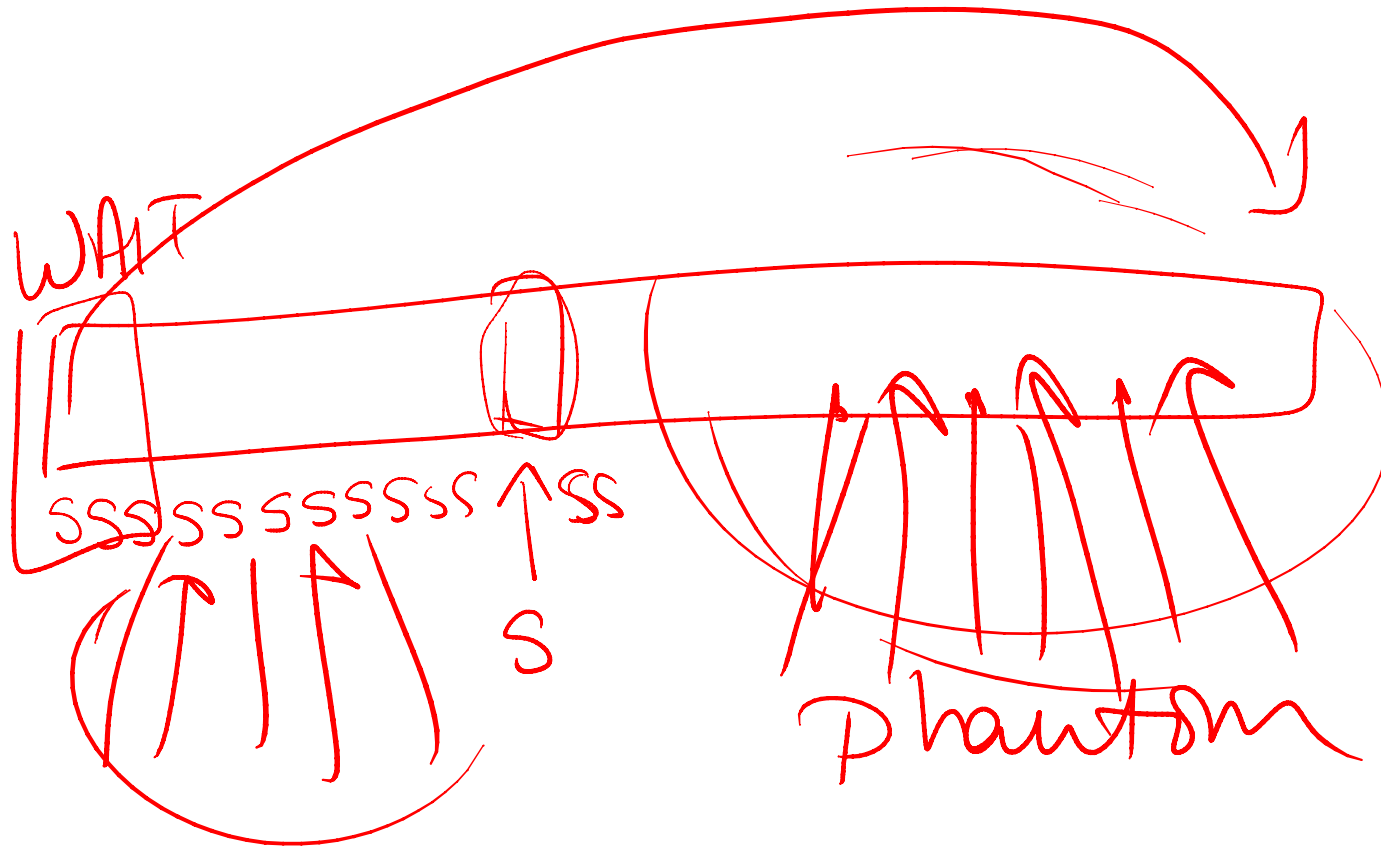
This is showing how an update to a CL key value that causes record relocation (after another query has already read [and released the lock on] the row) allows the reader to read the row TWICE (non-repeatedly).

This is from the demo on non-repeatable reads and shows what I called the “anderson-zembrowsky” problem. 😊

Discussion: Phantoms (Inconsistent Analysis)

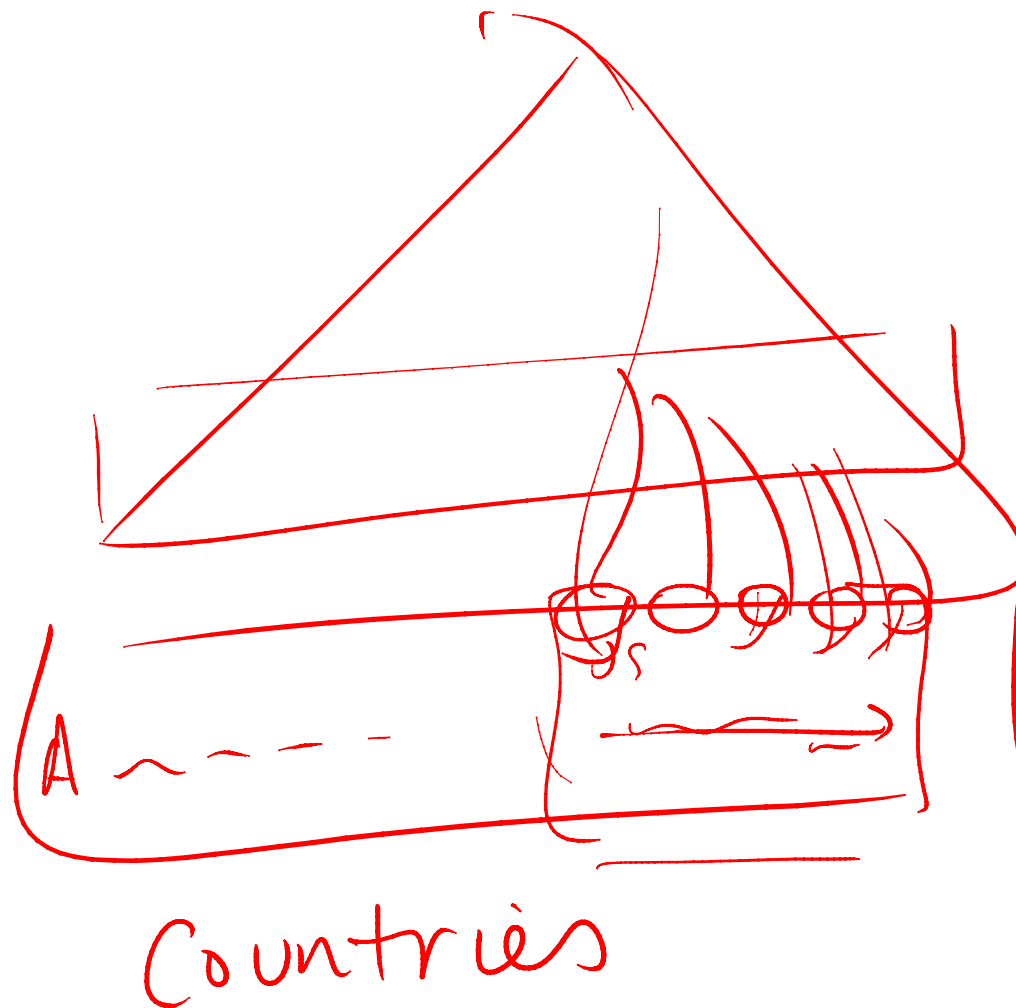
What is correct? Are we talking about row-consistency or statement-level consistency? The standards only define the state of the row at the time that the row is accessed. There is NOTHING that ties together the state of those rows at the time the query is accessed (unless you use versioning). So, here I was discussing the fact that NEW rows that are coming into the set will be visible in all isolation levels (except serializable). I always think of this scenario like a dog chasing its tail. 😊





Discussion: How does repeatable read prevent the Anderson/Zembrowsky problem?

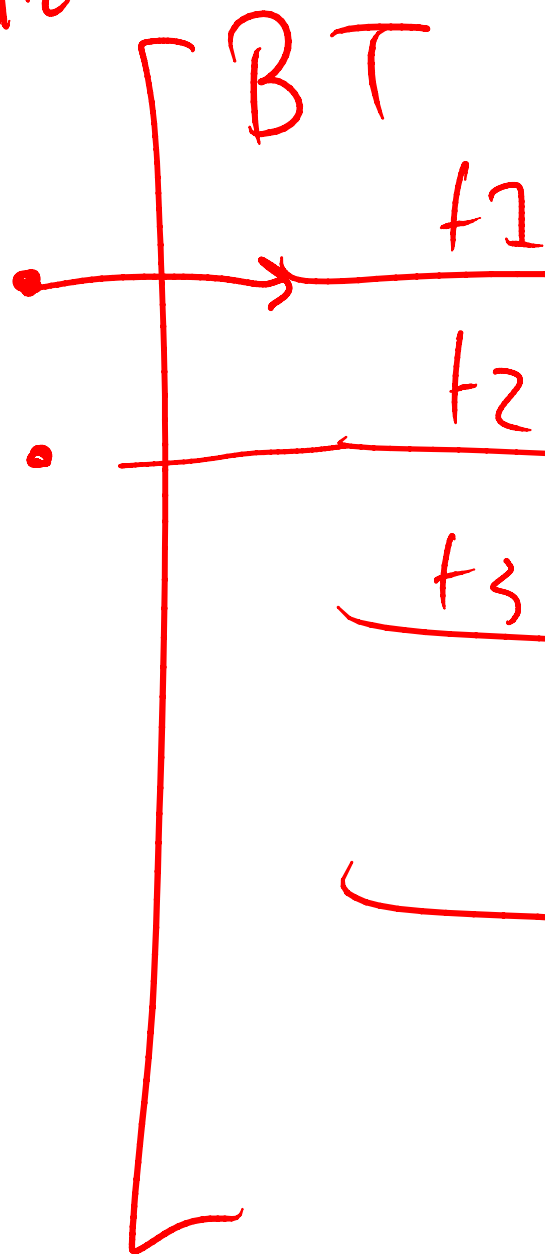
As rows are read, the shared locks used to read them are left behind to protect the rows. The rows cannot be read again in any other state because the shared locks will prevent other users from modifying the rows.



S11 – Serializable transactions

This picture described key range locking in an index. When a query runs the isolation level dictates the state of the data (even in the bounds of a transaction) that we should see. In a serializable transaction the data is isolated at the time the statement runs. As a result, data must be locked at that time. To prevent new rows from coming into the set, SQL Server will lock the “range” of rows affected by the query. If a good index exists then SQL Server can lock within the index (it’s in the first intermediate level – the level above the leaf level). If a good index does not exist then SQL Server must do table-level locking.

Serializ-

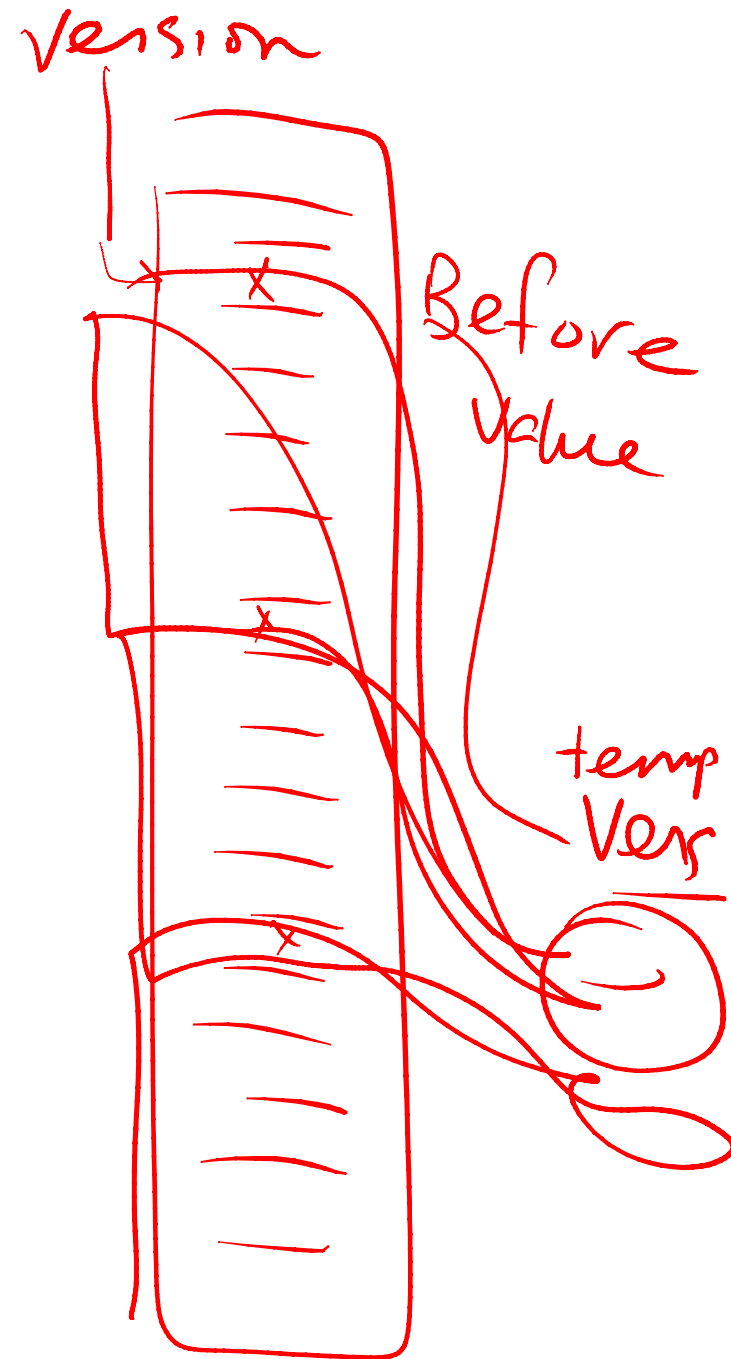
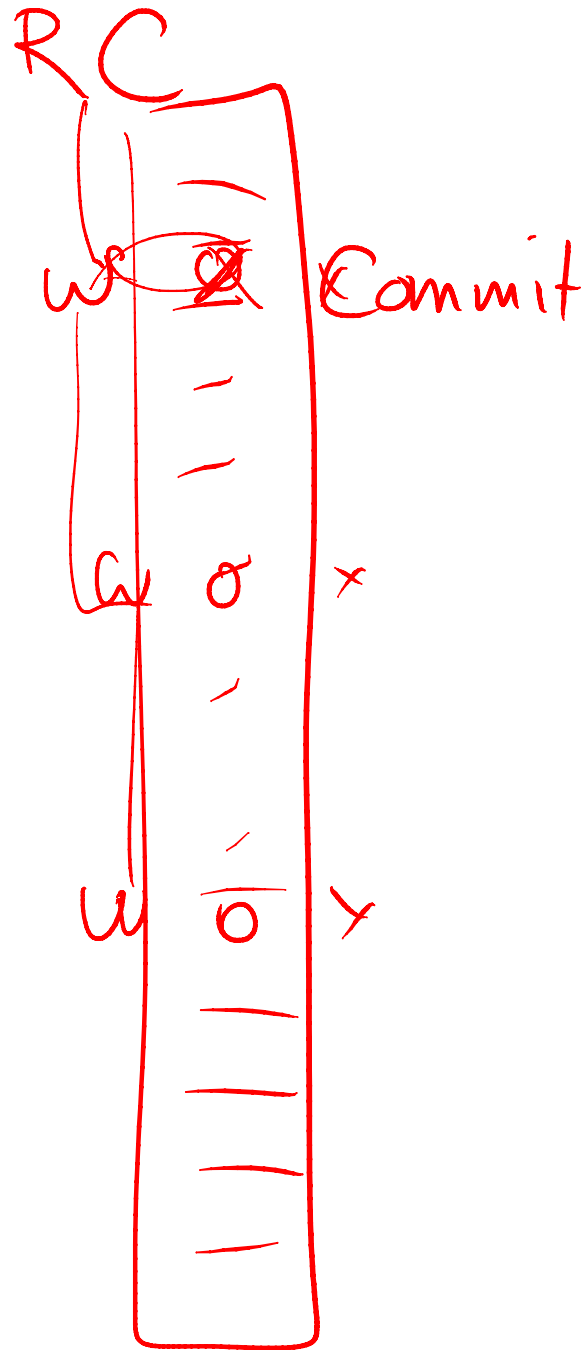
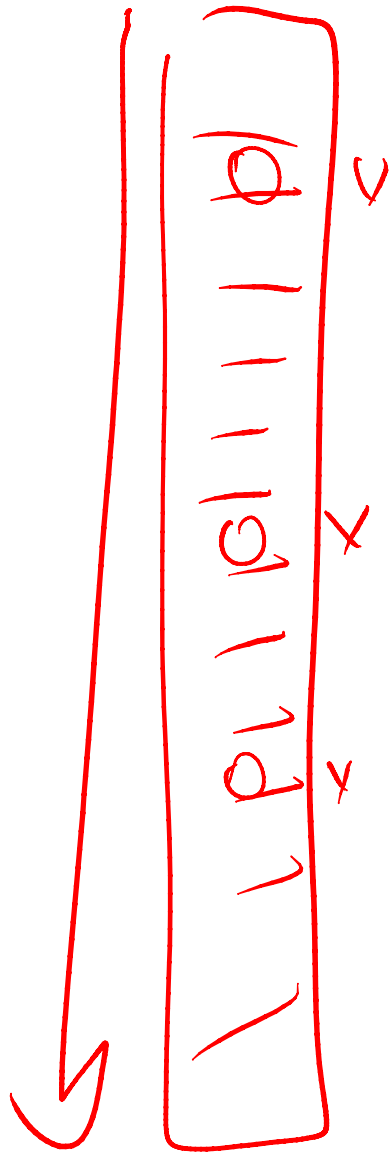


S20 – Snapshot Isolation vs. Serializable transactions

Sometimes I like to say that S.I. is more serializable than the locking based serializable transaction isolation level is. The reason why – S.I. defines the point to which all statements reconcile as the BEGINNING of the transaction. All statements in the transaction reconcile to EXACTLY that same point in time (defined by an LSN). However, serializable can only lock resources AS it requests them. In this picture I drew that distinction – each statement locks AS it executes. So, the 2nd and 3rd statements reconcile to t2 (time 2) and t3 (time 3). Whereas a S.I. transaction would always reconcile to t1.

See slide 24 for more information on transaction-level read consistency = Snapshot Isolation (using versioning).

⓪ nolock



See notes on next page

The prior slide showed a table vertically as a set of pages. The idea was to compare/contrast the behaviors between:

(1) The default – READ COMMITTED (using locking)

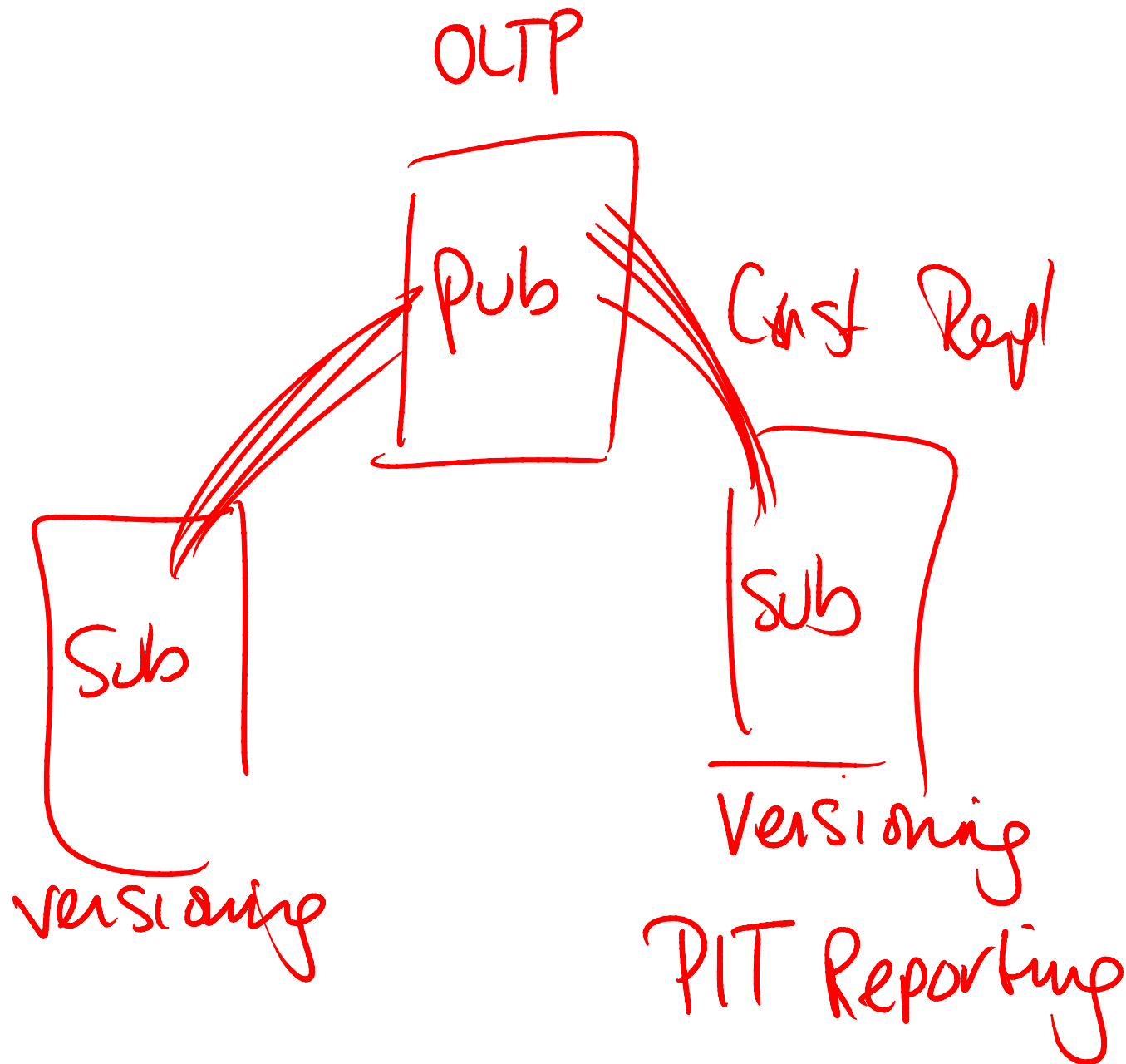
The key things to remember from the picture is that (1) has “hiccups” along the way as they encounter rows that are locked. They read... wait... read... wait. This is *partially* what slows down a statement. However, it’s worse than this. It’s possible that AFTER a row is read, it will appear again (if the record is relocated) and we will read it twice (non-repeatable reads). Also, it’s possible that another transaction would modify a row that we’ve NOT yet seen (before we get there) as well as a row that we have seen (after we’ve read it) such that the end result of our statement has rows that aren’t really transactionally consistent (with another multi-statement transaction). This is also a problem... The default (1) environment is prone to a few inconsistencies (aka. Inconsistent analysis). Only way to solve – increase isolation (or [as of 2005] consider using versioning).

(2) A forced NOLOCK

The key thing to remember here is that no lock allows the reader to read quickly – not stopping for locked rows. However, these rows may be “in-flight” and their modified data might end up getting rolled back or even be in a mid-flight state. If you’re looking for only an estimate – this might be fine.

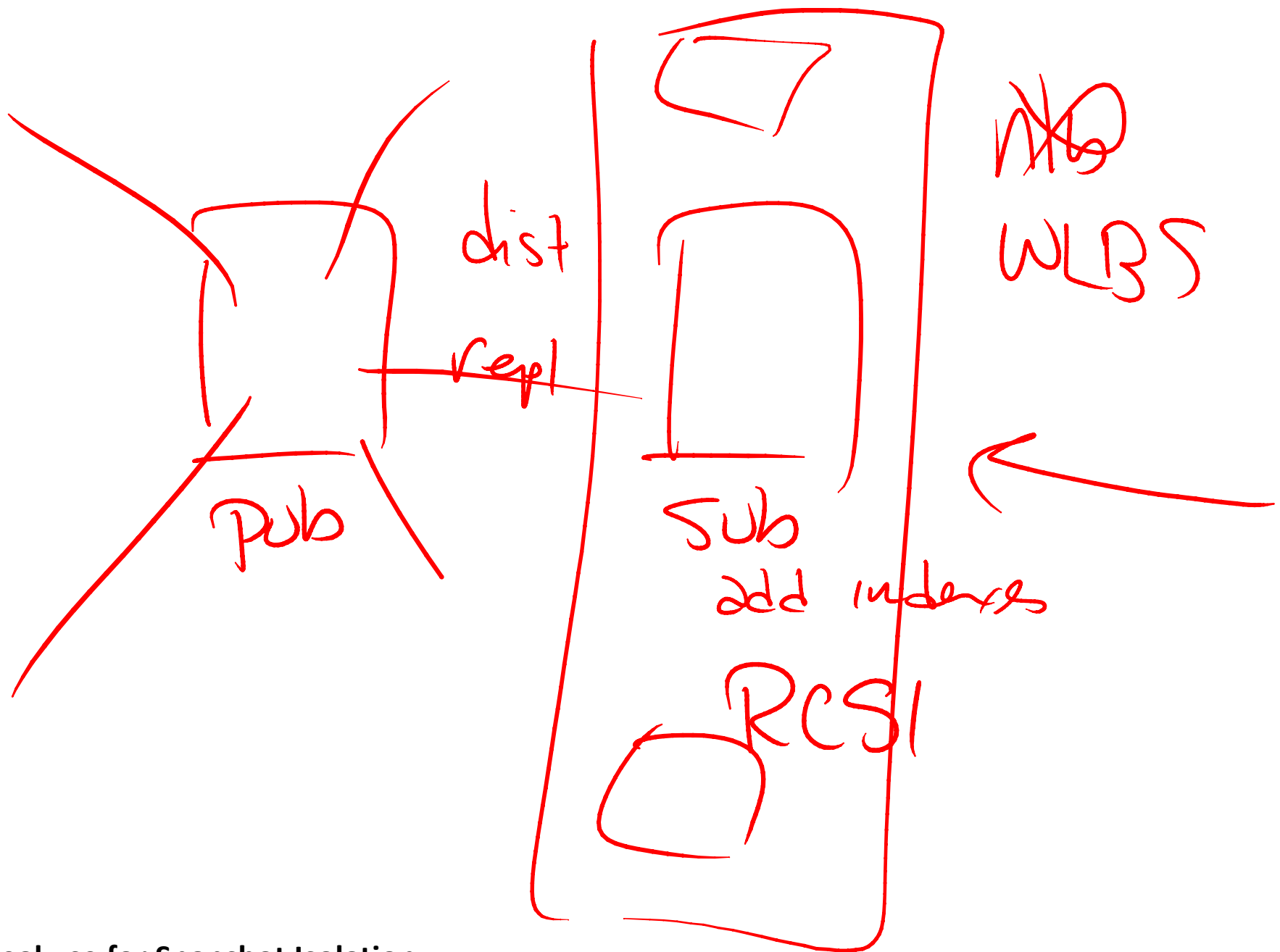
(3) Versioning (and it was implied that it was statement-level)

When a versioned query runs the reader will not stop but will have to go to the version store for any row that’s in flight. This is a tiny bit slower but guarantees consistency of the ENTIRE read to the point in time when the statement started. This gives you better concurrency as well as a definable point in time to which your statement reconciles. Of course, it isn’t free – the overhead of versioning is for every writer and it occurs within tempdb.



An ideal use for Snapshot Isolation

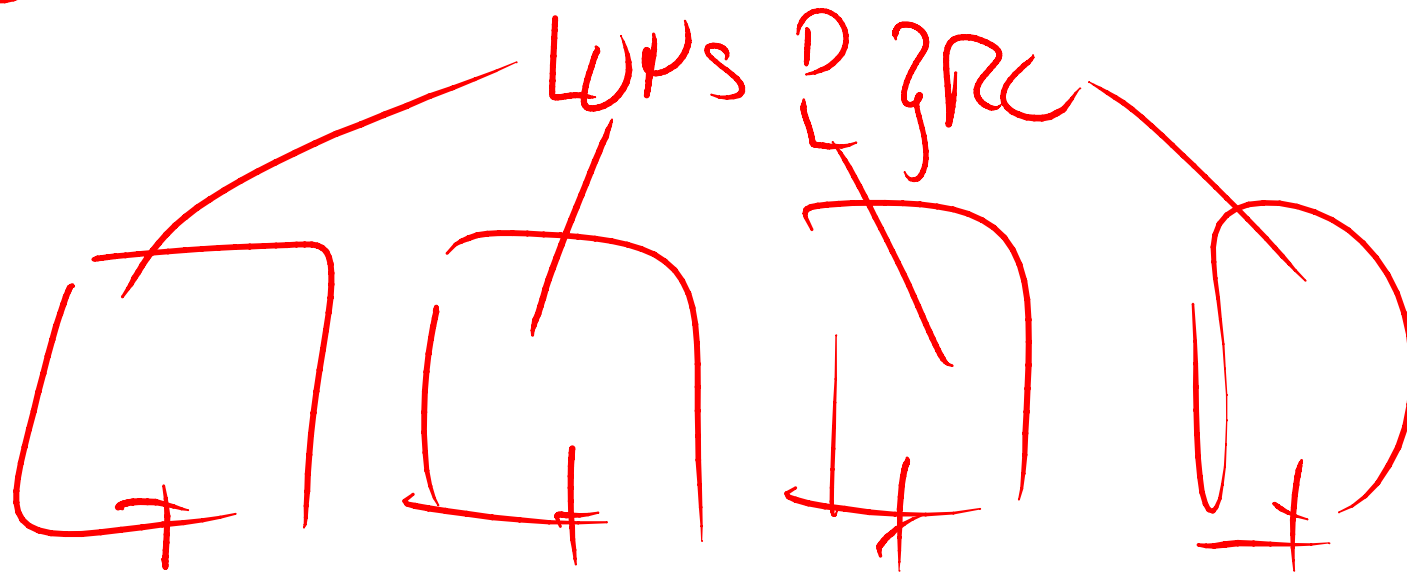
The idea is that you don't want to impact your production OLTP environment with READERS or with versioning so instead, you replicate to subscribers and use read committed snapshot there!



An ideal use for Snapshot Isolation

Taking that same idea a bit further – you could use WLBS (windows load balancing services) to create a server federation and let windows to load balancing by virtualizing the IP. But, to update these you'd have to have dual NICs and a bit more complicated setup.

SSD = Scalable Shared Databases



~~LUNs~~

For a SINGLE copy of a database on LUNs set to Read Only

You can create a federation using SSD (Scalable Shared Databases). Check out the BOL:

[http://technet.microsoft.com/en-us/library/ms345392\(v=sql.105\).aspx](http://technet.microsoft.com/en-us/library/ms345392(v=sql.105).aspx)