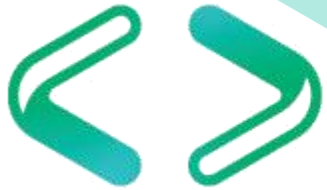
Let's Explore SQL Storage Internals

Brian Hansen
brian@tf3604.com
@tf3604



PASS SQLSATURDAY
LINCOLN | OCT 27 2018

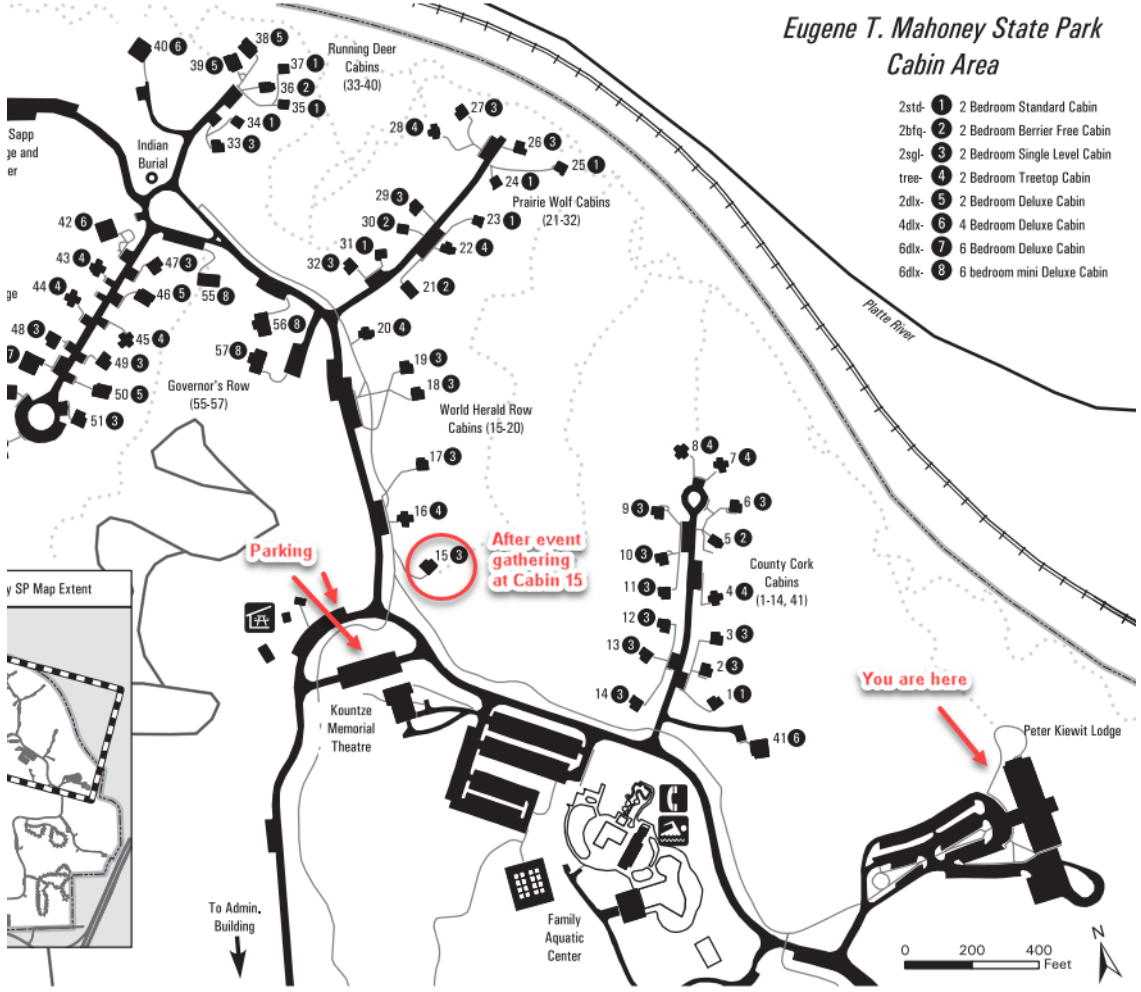
Welcome to SQLSaturday #767!
Hosted by
Lincoln SQL Server User Group

Visit http://Lincoln.pass.org for meeting & group information

Be sure to give our sponsors some #SQLLove

After Event Networking Gathering

Stop over to Cabin #15 after the event to network with #SQLFamily



Brian Hansen


brian@tf3604.com


@tf3604.com

children.org

- 20 Years working with SQL Server
  - Development work since 7.0
  - Administration going back to 6.5
  - Fascinated with SQL internals

www.tf3604.com/internals

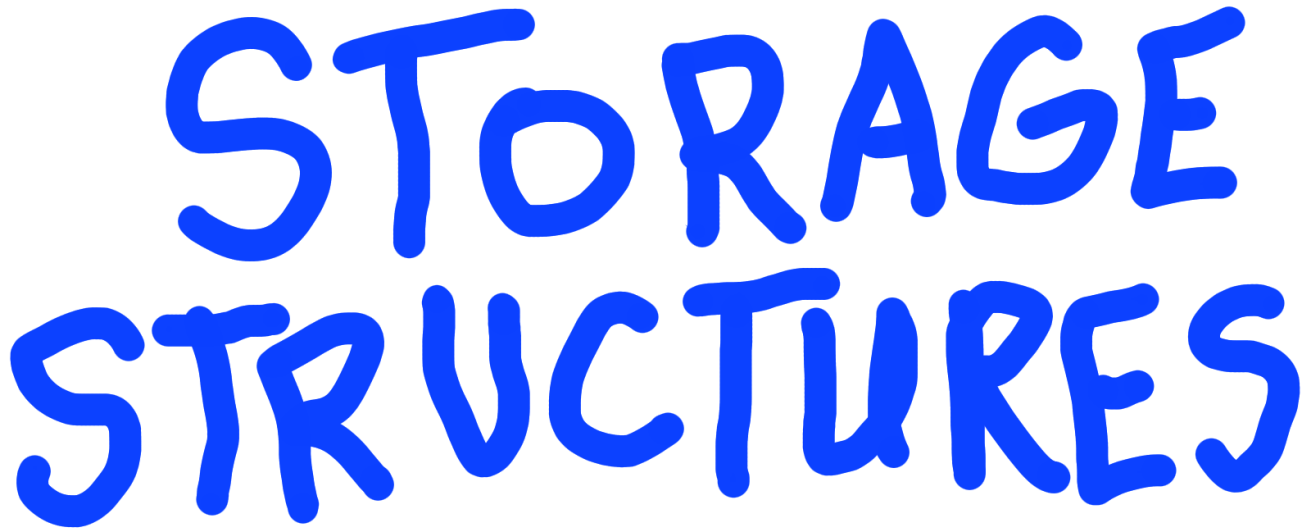
Agenda

- Why understand internals
- Basic data file structures
- GAM, SGAM, PFS and IAM pages
- Other page structures
- Data and index pages

Why understand storage internals?

- Stronger foundation for
  - Designing databases and tables
    - Maximizing storage utilization
    - Better performance
  - Help the optimizer come up with good plans
    - Better performance
- Better performance

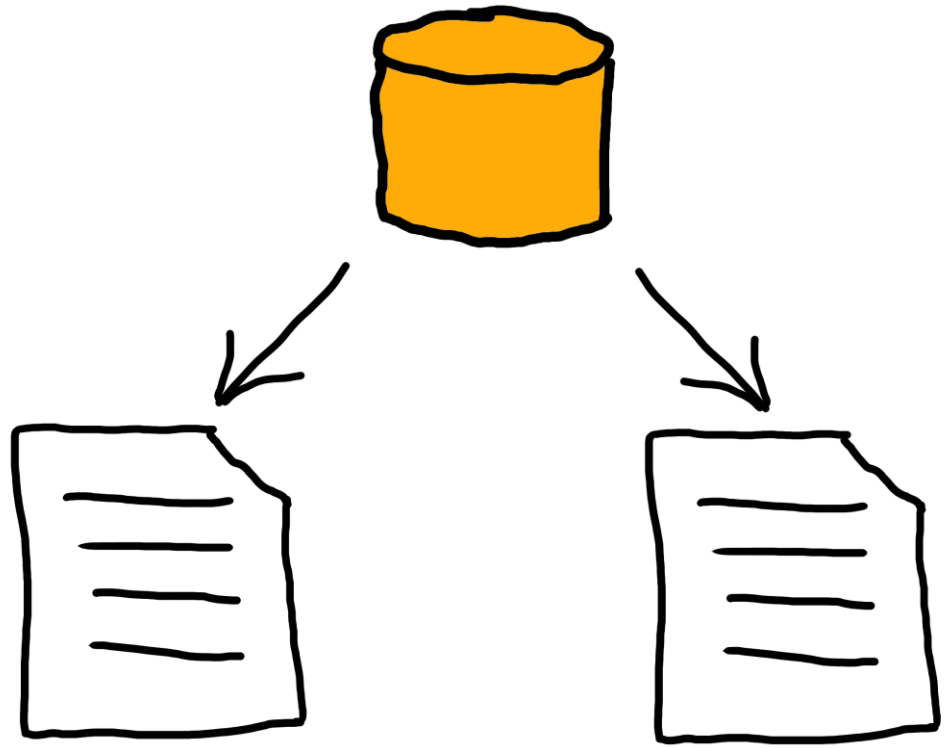
Prologue



This thing we call a database …

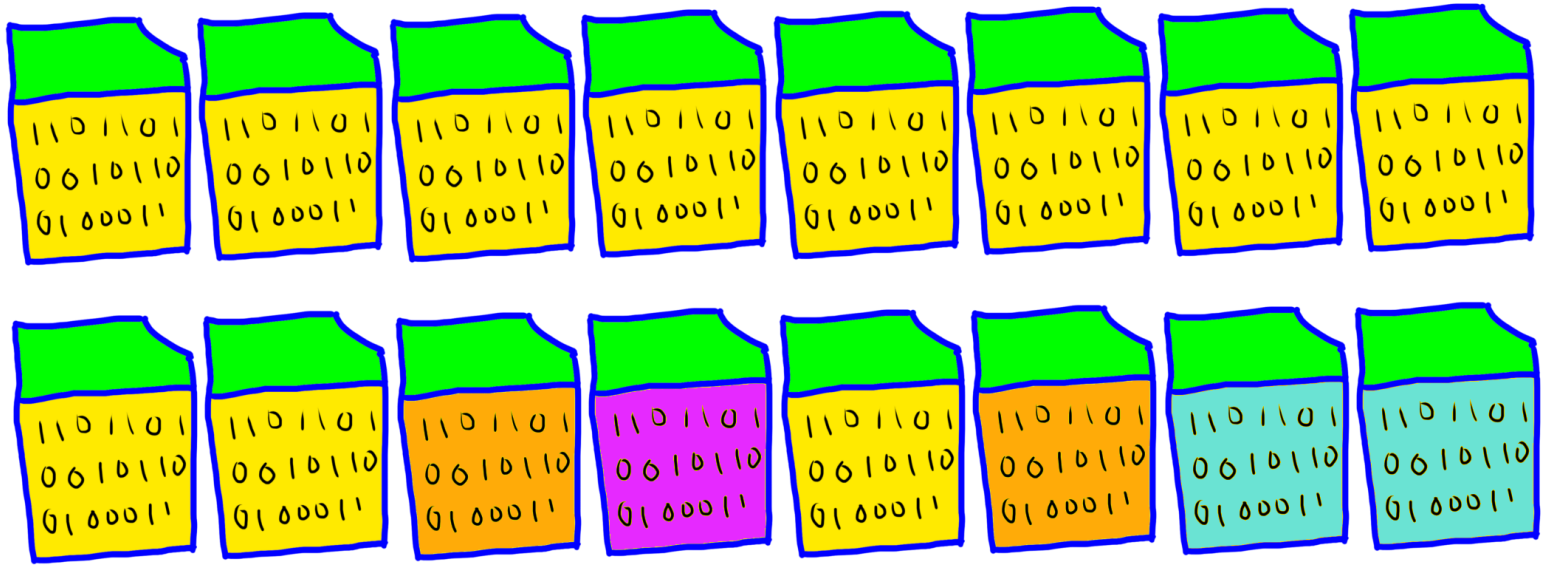
… is really just a couple of files *

- Data file
- Log file



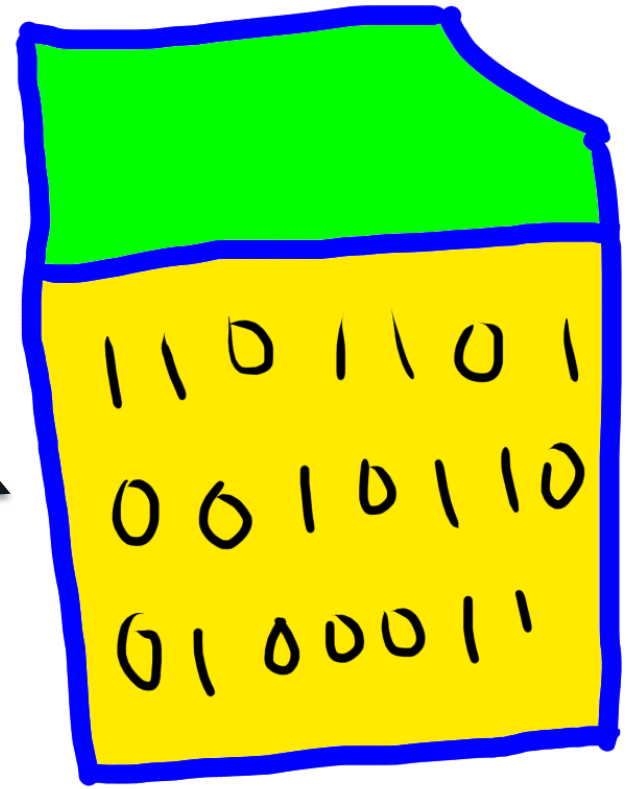
The data file

- Basic unit is the 8KB page
- Grouped into extents of 8 pages (64KB)
  - Extents can be “mixed” or “uniform”



Pages

- Again, are 8KB in size (8192 bytes)
  - 96-byte header
  - 8096-byte payload



What goodies are the page header?

- Page number (starts at 0)
  - Reported as (1:1234) ➔ file 1, page 1234
- Page type (details on next slide)
- Object ID of page's owner
- Allocation Unit ID of page's owner
- Last update LSN of page
- Checksum
- Pointer to previous and next page in chain
- And much more

Page types

ID	Description
1	Data
2	Index
3	Mixed text
4	Text data
8	GAM
9	SGAM
10	IAM
11	PFS
13	Boot

ID	Description
15	File header
16	DCM
17	BCM
7	Intermediate (sort)
18	Intermediate (CHECKDB)
19	Intermediate (reorg)
20	Intermediate (bulk load)

Types of tables

- Heaps
  - Data is not ordered
- Clustered table
  - Clustered index defines table order
  - Data structure contains multiple “levels”
    - Root node (page)
    - Internal nodes
    - Leaf nodes

Part 1

BITMAPS

GAM pages (global allocation map)

- Page 2 in file
- Page (511,232 ✕ n)† in file (every 3.90 GB)
- Each bit in the page tracks one extent in the file

Value	Meaning
0	Extent is allocated
1	Extent is not allocated

† Unless (511,232 ✕ n) is a multiple of 8088, in which case the GAM page falls on (511,232 ✕ n) + 1

SGAM pages (shared global allocation map)

- Page 3 in file
- Page (511,232 ✕ n + 1)† in file (every 3.90 GB)
- Each bit in the page tracks one extent in the file

Value	Meaning
0	Dedicated extent or mixed extent that is full
1	Mixed extent with unallocated pages

† Unless (511,232 ✕ n) is a multiple of 8088, in which case the SGAM page falls on (511,232 ✕ n) + 2

Putting GAM and SGAM together

GAM	Meaning
0	Extent is allocated
1	Extent is not allocated

SGAM	Meaning
0	Dedicated extent or mixed extent (full)
1	Mixed extent with unallocated pages

GAM bit	SGAM bit	Meaning
0	0	Dedicated extent or mixed extent that is full
0	1	Mixed extent with unallocated pages
1	0	Free extent, not in use
1	1	**ERROR: Invalid**

IAM pages (index allocation map)

- Associated to an allocation unit
- Each bit tracks one extent in the file
- But wait … what is an “allocation unit”?

Allocation units

- Introduced in SQL Server 2005
- Can be one of the following:
  - Hobt (heap or b-tree)
  - LOB data
  - Row overflow data (SLOB)
- See sys.allocation_units

Partitions

- Every table contains one or more partitions
  - Multiple partitions only supported in Enterprise Edition up through SQL 2016 SP1
  - Supported in all editions starting with SQL 2016 SP1
- Non-clustered indexes on a partitioned table will also have multiple partitions

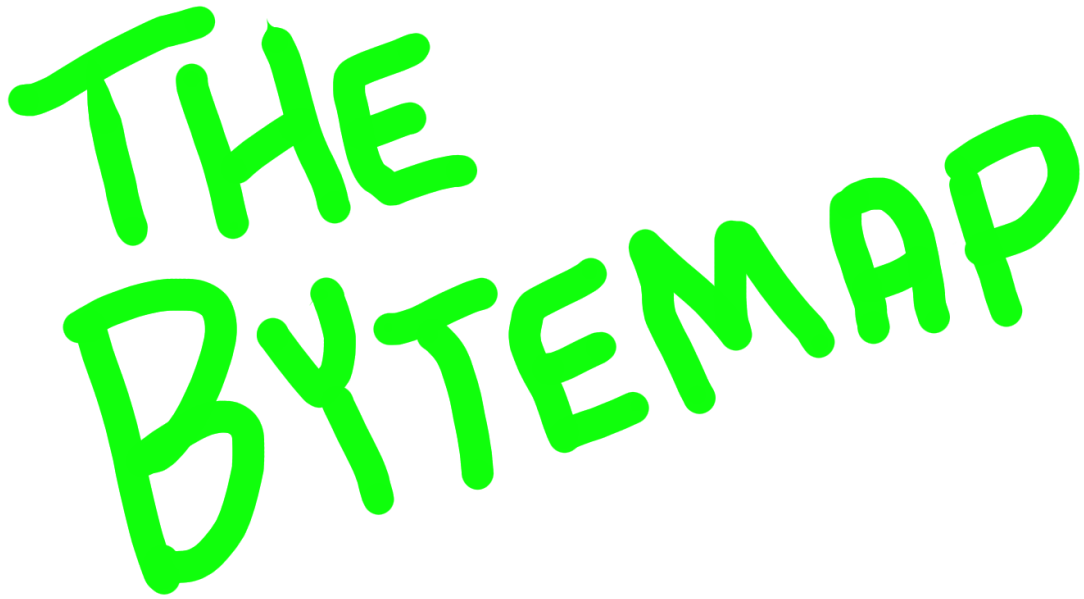
Back to IAM pages

- Associated to an allocation unit
- Each bit tracks one extent in the file
- So a table will have one IAM page for each
  - Partition
    - Index (clustered / heap and non-clustered)
      - In-row data
      - LOB data
      - Overflow data
- Each of these is the first in an “IAM chain” of pages

What does the IAM bit mean?

Value	Meaning
0	Extent is allocated to the allocation unit
1	Extent is not allocated to the allocation unit

Part 2



PFS pages (page free space)

- Page 1 in file
- Page (8,088 ✕ n) in file (every 63 MB)
- Each byte in the page tracks one page in the file
- Different bits in the byte have specific meanings

PFS pages (page free space)

Bit #	Description	Value	Meaning
0 to 2	Percent of free space on LOB or heap page	0	Page is empty
		1	Page is 1% to 50% full
		2	Page is 51% to 80% full
		3	Page is 81% to 95% full
		4	Page is >= 96% full
3	Ghost records	0	Page has no ghost records
		1	Page has ghost records

PFS pages (page free space)

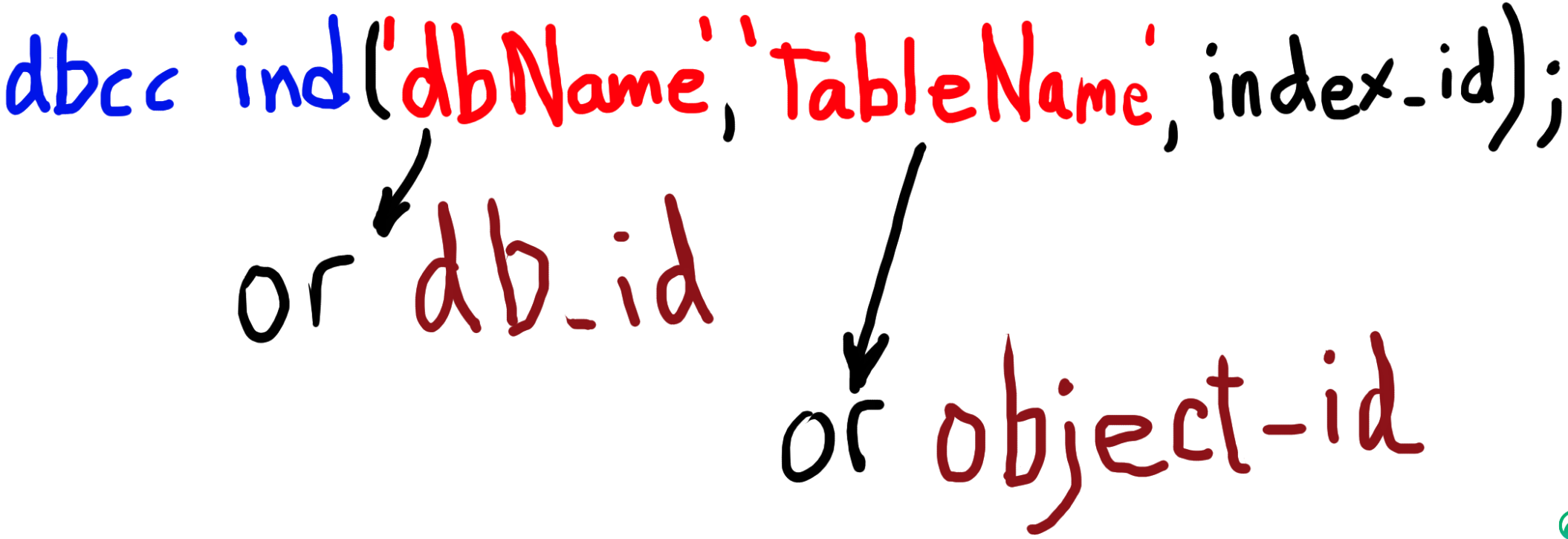
Bit #	Description	Value	Meaning
4	IAM page	0	Page is not an IAM page
		1	Page is an IAM page
5	Mixed page	0	Page is not on a mixed extent
		1	Page is on a mixed extent
6	Allocation	0	Page is not allocated
		1	Page is allocated
7	Not used		

Part 3

VIEWING PAGES

DBCC IND

- Undocumented
- Returns the IAM chains associated to an object



DBCC IND example

```
dbcc ind ('CorpDB', 'Customer', 1);
```

DBCC IND

	PageFID	PagePID	IAMFID	IAMPID	ObjectID	IndexID	PartitionNumber	PartitionID	iam_chain_type	PageType	IndexLevel	NextPageFID	NextPagePID	PrevPageFID	PrevPagePID
1	1	489	NULL	NULL	245575913	1	1	72057594039042048	In-row data	10	NULL	0	0	0	0
2	1	490	1	489	245575913	1	1	72057594039042048	In-row data	2	1	1	624	0	0
3	1	513	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	514	1	608
4	1	514	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	515	1	513
5	1	515	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	516	1	514
6	1	516	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	517	1	515
7	1	517	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	518	1	516
8	1	518	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	519	1	517
9	1	519	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	609	1	518
10	1	417	1	489	245575913	1	1	72057594039042048	In-row data	1	0	0	0	1	1695
11	1	455	1	489	245575913	1	1	72057594039042048	In-row data	2	2	0	0	0	0
12	1	528	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	529	0	0
13	1	529	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	530	1	528
14	1	530	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	531	1	529
15	1	531	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	532	1	530
16	1	532	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	533	1	531
17	1	533	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	534	1	532
18	1	534	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	535	1	533
19	1	535	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	536	1	534
20	1	536	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	537	1	535
21	1	537	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	538	1	536
22	1	538	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	539	1	537
23	1	539	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	540	1	538
24	1	540	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	541	1	539
25	1	541	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	542	1	540
26	1	542	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	543	1	541
27	1	543	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	544	1	542
28	1	544	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	545	1	543
29	1	545	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	546	1	544

DBCC IND

	PageFID	PagePID	IAMFID	IAMPID	ObjectID	IndexID	PartitionNumber	PartitionID	iam_chain_type	PageType	IndexLevel	NextPageFID	NextPagePID	PrevPageFID	PrevPagePID
1	1	489	NULL	NULL	245575913	1	1	72057594039042048	In-row data	10	NULL	0	0	0	0
2	1	490	1	489	245575913	1	1	72057594039042048	In-row data	2	1	1	624	0	0
3	1	513	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	514	1	608
4	1	514	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	515	1	513
5	1	515	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	516	1	514
6	1	516	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	517	1	515
7	1	517	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	518	1	516
8	1	518	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	519	1	517
9	1	519	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	609	1	518
10	1	417	1	489	245575913	1	1	72057594039042048	In-row data	1	0	0	0	1	1695
11	1	455	1	489	245575913	1	1	72057594039042048	In-row data	2	2	0	0	0	0
12	1	528	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	529	0	0
13	1	529	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	530	1	528
14	1	530	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	531	1	529
15	1	531	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	532	1	530
16	1	532	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	533	1	531
17	1	533	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	534	1	532
18	1	534	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	535	1	533
19	1	535	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	536	1	534
20	1	536	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	537	1	535
21	1	537	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	538	1	536
22	1	538	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	539	1	537
23	1	539	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	540	1	538
24	1	540	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	541	1	539
25	1	541	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	542	1	540
26	1	542	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	543	1	541
27	1	543	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	544	1	542
28	1	544	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	545	1	543
29	1	[illegible]	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	546	1	544

DBCC IND

	PageFID	PagePID	IAMFID	IAMPID	ObjectID	IndexID	PartitionNumber	PartitionID	iam_chain_type	PageType	IndexLevel	NextPageFID	NextPagePID	PrevPageFID	PrevPagePID
1	1	489	NULL	NULL	245575913	1	1	72057594039042048	In-row data	10	NULL	0	0	0	0
2	1	490	1	489	245575913	1	1	72057594039042048	In-row data	2	1	1	624	0	0
3	1	513	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	514	1	608
4	1	514	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	515	1	513
5	1	515	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	516	1	514
6	1	516	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	517	1	515
7	1	517	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	518	1	516
8	1	518	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	519	1	517
9	1	519	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	609	1	518
10	1	417	1	489	245575913	1	1	72057594039042048	In-row data	1	0	0	0	1	1695
11	1	455	1	489	245575913	1	1	72057594039042048	In-row data	2	2	0	0	0	0
12	1	528	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	529	0	0
13	1	529	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	530	1	528
14	1	530	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	531	1	529
15	1	531	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	532	1	530
16	1	532	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	533	1	531
17	1	533	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	534	1	532
18	1	534	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	535	1	533
19	1	535	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	536	1	534
20	1	536	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	537	1	535
21	1	537	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	538	1	536
22	1	538	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	539	1	537
23	1	539	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	540	1	538
24	1	540	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	541	1	539
25	1	541	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	542	1	540
26	1	542	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	543	1	541
27	1	543	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	544	1	542
28	1	544	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	545	1	543
29	1	545	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	546	1	544

DBCC IND

	PageFID	PagePID	IAMFID	IAMPID	ObjectID	IndexID	PartitionNumber	PartitionID	iam_chain_type	PageType	IndexLevel	NextPageFID	NextPagePID	PrevPageFID	PrevPagePID
1	1	489	NULL	NULL	245575913	1	1	72057594039042048	In-row data	10	NULL	0	0	0	0
2	1	490	1	489	245575913	1	1	72057594039042048	In-row data	2	1	1	624	0	0
3	1	513	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	514	1	608
4	1	514	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	515	1	513
5	1	515	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	516	1	514
6	1	516	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	517	1	515
7	1	517	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	518	1	516
8	1	518	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	519	1	517
9	1	519	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	609	1	518
10	1	417	1	489	245575913	1	1	72057594039042048	In-row data	1	0	0	0	1	1695
11	1	455	1	489	245575913	1	1	72057594039042048	In-row data	2	2	0	0	0	0
12	1	528	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	529	0	0
13	1	529	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	530	1	528
14	1	530	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	531	1	529
15	1	531	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	532	1	530
16	1	532	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	533	1	531
17	1	533	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	534	1	532
18	1	534	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	535	1	533
19	1	535	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	536	1	534
20	1	536	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	537	1	535
21	1	537	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	538	1	536
22	1	538	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	539	1	537
23	1	539	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	540	1	538
24	1	540	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	541	1	539
25	1	541	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	542	1	540
26	1	542	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	543	1	541
27	1	543	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	544	1	542
28	1	544	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	545	1	543
29	1	545	1	489	[illegible]	1	1	[illegible]	In-row data	1	0	1	546	1	544

DBCC IND

	PageFID	PagePID	IAMFID	IAMPID	ObjectID	IndexID	PartitionNumber	PartitionID	iam_chain_type	PageType	IndexLevel	NextPageFID	NextPagePID	PrevPageFID	PrevPagePID
1	1	489	NULL	NULL	245575913	1	1	72057594039042048	In-row data	10	NULL	0	0	0	0
2	1	490	1	489	245575913	1	1	72057594039042048	In-row data	2	1	1	624	0	0
3	1	513	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	514	1	608
4	1	514	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	515	1	513
5	1	515	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	516	1	514
6	1	516	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	517	1	515
7	1	517	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	518	1	516
8	1	518	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	519	1	517
9	1	519	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	609	1	518
10	1	417	1	489	245575913	1	1	72057594039042048	In-row data	1	0	0	0	1	1695
11	1	455	1	489	245575913	1	1	72057594039042048	In-row data	2	2	0	0	0	0
12	1	528	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	529	0	0
13	1	529	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	530	1	528
14	1	530	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	531	1	529
15	1	531	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	532	1	530
16	1	532	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	533	1	531
17	1	533	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	534	1	532
18	1	534	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	535	1	533
19	1	535	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	536	1	534
20	1	536	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	537	1	535
21	1	537	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	538	1	536
22	1	538	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	539	1	537
23	1	539	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	540	1	538
24	1	540	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	541	1	539
25	1	541	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	542	1	540
26	1	542	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	543	1	541
27	1	543	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	544	1	542
28	1	544	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	545	1	543
29	1	545	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	546	1	544

DBCC IND

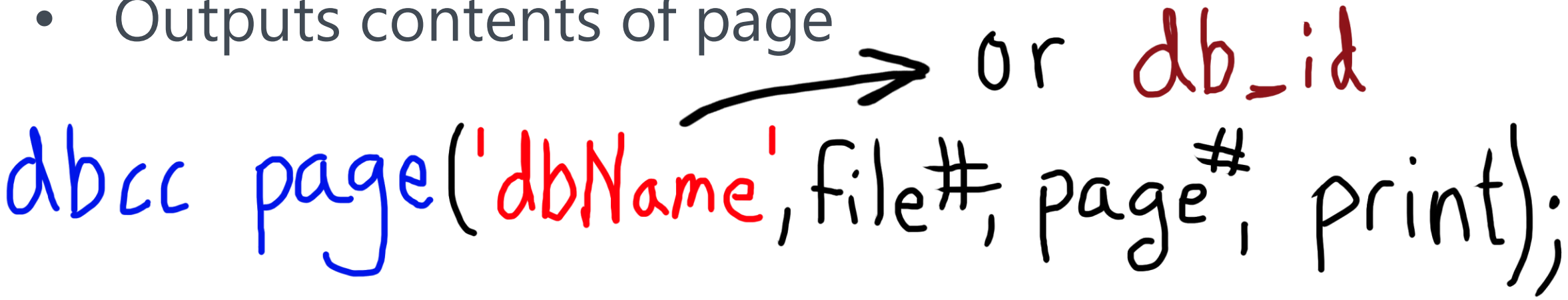
	PageFID	PagePID	IAMFID	IAMPID	ObjectID	IndexID	PartitionNumber	PartitionID	iam_chain_type	PageType	IndexLevel	NextPageFID	NextPagePID	PrevPageFID	PrevPagePID
1	1	489	NULL	NULL	245575913	1	1	72057594039042048	In-row data	10	NULL	0	0	0	0
2	1	490	1	489	245575913	1	1	72057594039042048	In-row data	2	1	1	624	0	0
3	1	513	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	514	1	608
4	1	514	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	515	1	513
5	1	515	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	516	1	514
6	1	516	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	517	1	515
7	1	517	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	518	1	516
8	1	518	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	519	1	517
9	1	519	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	609	1	518
10	1	417	1	489	245575913	1	1	72057594039042048	In-row data	1	0	0	0	1	1695
11	1	455	1	489	245575913	1	1	72057594039042048	In-row data	2	2	0	0	0	0
12	1	528	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	529	0	0
13	1	529	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	530	1	528
14	1	530	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	531	1	529
15	1	531	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	532	1	530
16	1	532	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	533	1	531
17	1	533	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	534	1	532
18	1	534	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	535	1	533
19	1	535	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	536	1	534
20	1	536	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	537	1	535
21	1	537	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	538	1	536
22	1	538	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	539	1	537
23	1	539	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	540	1	538
24	1	540	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	541	1	539
25	1	541	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	542	1	540
26	1	542	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	543	1	541
27	1	543	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	544	1	542
28	1	544	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	545	1	543
29	1	545	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	546	1	544

DBCC IND

	PageFID	PagePID	IAMFID	IAMPID	ObjectID	IndexID	PartitionNumber	PartitionID	iam_chain_type	PageType	IndexLevel	NextPageFID	NextPagePID	PrevPageFID	PrevPageID
1	1	489	NULL	NULL	245575913	1	1	72057594039042048	In-row data	10	NULL	0	0	0	0
2	1	490	1	489	245575913	1	1	72057594039042048	In-row data	2	1	1	624	0	0
3	1	513	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	514	1	608
4	1	514	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	515	1	513
5	1	515	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	516	1	514
6	1	516	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	517	1	515
7	1	517	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	518	1	516
8	1	518	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	519	1	517
9	1	519	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	609	1	518
10	1	417	1	489	245575913	1	1	72057594039042048	In-row data	1	0	0	0	1	1695
11	1	455	1	489	245575913	1	1	72057594039042048	In-row data	2	2	0	0	0	0
12	1	528	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	529	0	0
13	1	529	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	530	1	528
14	1	530	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	531	1	529
15	1	531	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	532	1	530
16	1	532	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	533	1	531
17	1	533	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	534	1	532
18	1	534	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	535	1	533
19	1	535	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	536	1	534
20	1	536	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	537	1	535
21	1	537	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	538	1	536
22	1	538	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	539	1	537
23	1	539	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	540	1	538
24	1	540	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	541	1	539
25	1	541	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	542	1	540
26	1	542	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	543	1	541
27	1	543	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	544	1	542
28	1	544	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	545	1	543
29	1	545	1	489	245575913	1	1	72057594039042048	In-row data	1	0	1	546	1	544

DBCC PAGE

- Undocumented; combine with trace flag 3604
- Outputs contents of page



print	meaning
0	Header only
1	Rows and slots

print	meaning
2	Page and slots
3	Detailed interpretation

DBCC PAGE example

```
dbcc traceon (3604);
dbcc page ('CorpDB', 1, 489, 3);
```

DBCC PAGE output

```
PAGE: (1:489)


BUFFER:


BUF @0x000000F93EA3FB40

bpage = 0x000000F6A3044000          bhash = 0x0000000000000000          bpageno = (1:489)
bdbid = 7                           breferences = 0                     bcputicks = 0
bsampleCount = 0                    bUse1 = 32504                       bstat = 0x9
blog = 0x15a                        bnext = 0x0000000000000000          bDirtyContext = 0x0000000000000000
bstat2 = 0x0
```

DBCC PAGE output

```
PAGE HEADER:


Page @0x000000F6A3044000

m_pageId = (1:489)                    m_headerVersion = 1                   m_type = 10
m_typeFlagBits = 0x0                  m_level = 0                           m_flagBits = 0x200
m_objId (AllocUnitId.idObj) = 84      m_indexId (AllocUnitId.idInd) = 256
Metadata: AllocUnitId = 72057594043432960
Metadata: PartitionId = 72057594039042048                                    Metadata: IndexId = 1
Metadata: ObjectId = 245575913        m_prevPage = (0:0)                    m_nextPage = (0:0)
pminlen = 90                          m_slotCnt = 2                         m_freeCnt = 6
m_freeData = 8182                     m_reservedCnt = 0                     m_lsn = (284:3512:666)
m_xactReserved = 0                    m_xdesId = (0:0)                      m_ghostRecCnt = 0
m_tornBits = 2056427858               DB Frag ID = 1
```

DBCC PAGE output

```
Allocation Status

GAM (1:2) = ALLOCATED                SGAM (1:3) = NOT ALLOCATED
PFS (1:1) = 0x70 IAM_PG MIXED_EXT ALLOCATED   0_PCT_FULL                 DIFF (1:6) = NOT CHANGED
ML (1:7) = NOT MIN_LOGGED

IAM: Header @0x000001010D34A064 Slot 0, Offset 96

sequenceNumber = 0                   status = 0x0                        objectId = 0
indexId = 0                          page_count = 0                      start_pg = (1:0)
```

DBCC PAGE output

```
IAM: Extent Alloc Status Slot 1 @0x000001010D34A0C2

(1:0)        - (1:408)      = NOT ALLOCATED
(1:416)      -              =     ALLOCATED
(1:424)      - (1:440)      = NOT ALLOCATED
(1:448)      -              =     ALLOCATED
(1:456)      - (1:520)      = NOT ALLOCATED
(1:528)      - (1:608)      =     ALLOCATED
(1:616)      -              = NOT ALLOCATED
(1:624)      -              =     ALLOCATED
(1:632)      - (1:744)      = NOT ALLOCATED
(1:752)      - (1:760)      =     ALLOCATED
(1:768)      -              = NOT ALLOCATED
(1:776)      - (1:784)      =     ALLOCATED
(1:792)      -              = NOT ALLOCATED
(1:800)      - (1:808)      =     ALLOCATED
(1:816)      -              = NOT ALLOCATED
(1:824)      -              =     ALLOCATED
(1:832)      -              = NOT ALLOCATED
(1:840)      - (1:848)      =     ALLOCATED
(1:856)      - (1:912)      = NOT ALLOCATED
(1:920)      - (1:928)      =     ALLOCATED
(1:936)      -              = NOT ALLOCATED
(1:944)      - (1:992)      =     ALLOCATED
(1:1000)     -              = NOT ALLOCATED
```

Part 4

EVEN MORE BITMAPS

DCM pages (differential change map)

- Page 6 in file
- Page (511,232 ✕ n + 6) in file (every 3.90 GB)
- Each bit in the page tracks one extent in the file

Value	Meaning
0	Extent is unchanged since last full backup
1	Extent is changed since last full backup

BCM pages (bulk change map)

- Page 7 in file
- Page (511,232 ✕ n + 7) in file (every 3.90 GB)
- Each bit in the page tracks one extent in the file

Value	Meaning
0	Extent has minimally logged operations since last log backup
1	Extent has no minimally logged operations since last log backup

Part 5

HEADERS

File header page

- One per file, always page 0
- Selected contents
  - Logical name
  - File size and growth
  - LSNs and GUIDs

```
dbcc fileheader('dbName',file#);
```

Boot page

- One per database, always page 9 in file 1
- Selected contents
  - Database name / version
  - Last backup times / LSNs / GUIDs
  - Last CHECKDB time
  - Database configuration settings

dbcc dbinfo;

The file headers and boot page are critical

- If one of these pages gets corrupted …
- … There is no magical repair option
- ** RESTORE FROM BACKUP **

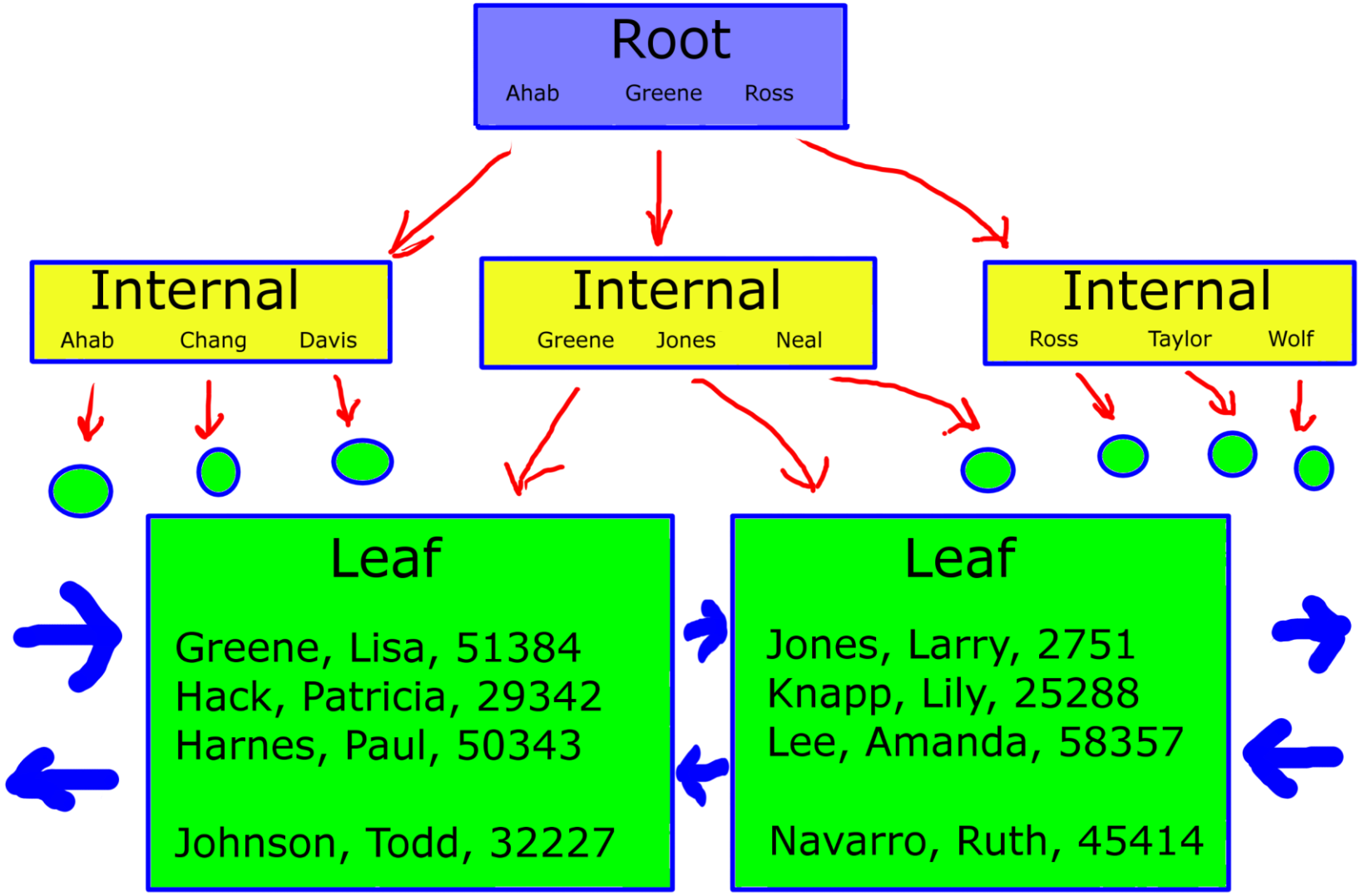
Part 6

Data Pages (the good stuff)

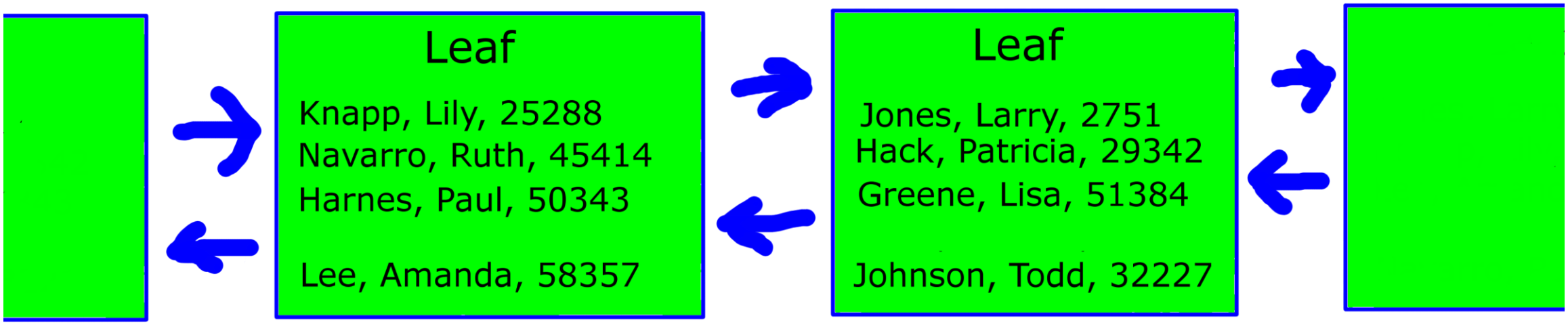
Types of data

- In-row data (data page)
- Index data (index page)
- LOB data (mixed text page or text data page)
- Other:
  - Forwarding records (heaps)
  - Ghost records

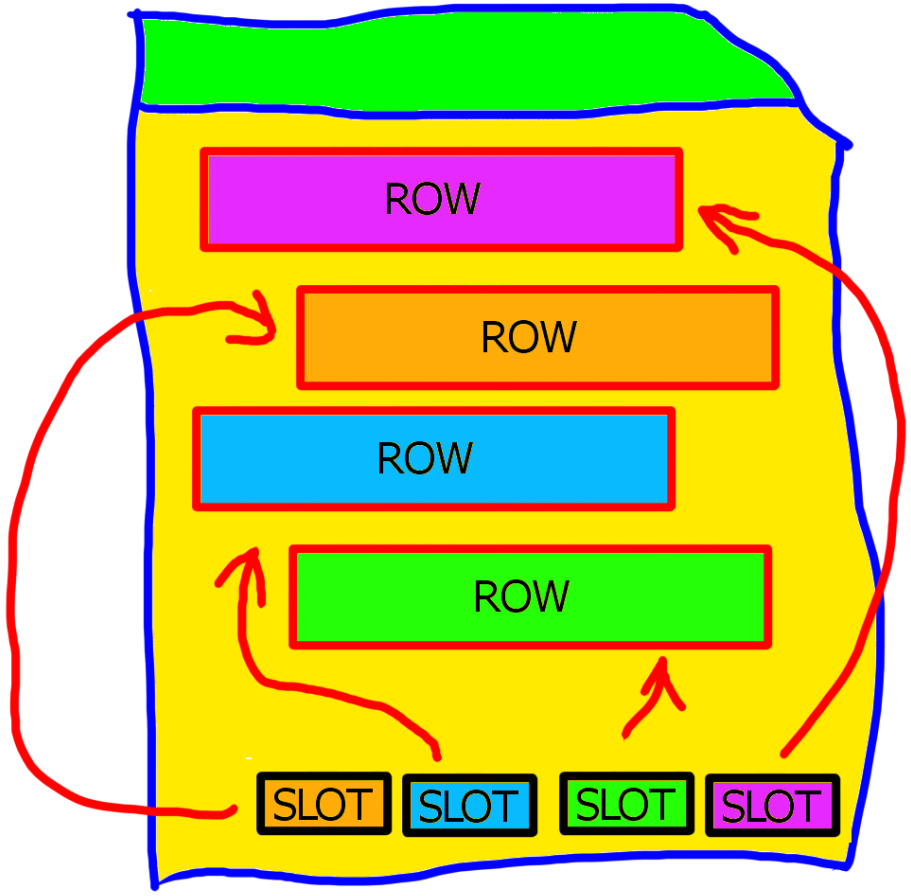
B-Tree structure



Heaps


Leaf
Knapp, Lily, 25288
Navarro, Ruth, 45414
Harnes, Paul, 50343
Lee, Amanda, 58357
Leaf
Jones, Larry, 2751
Hack, Patricia, 29342
Greene, Lisa, 51384
Johnson, Todd, 32227

Data pages



Record structure

- NULL bitmap
  - One bit per column (regardless of whether the column is nullable)
- Fixed-length data
  - int, bigint, char(x), nchar(x), binary(x), datetime, datetime2(x)
- Variable-length data
  - varchar(x), nvarchar(x), varbinary(x)

Record structure

Name	Size (bytes)	Description
Status	2	Status bits
FSize	2	Fixed-length data size
FData	FSize - 4	Fixed-length data
ColCount	2	Number of columns
NullBitMap	ColCount / 8	NULL bitmap (0 = not null; 1 = null)
VCount	2	Number of variable-length columns
VOffsets	VCount ✕ 2	Variable-length column offsets
VData	variable	Variable-length data

Data record example

```
create table dbo.CustomerInfo
(
   CustomerID int not null,
   FirstName varchar(50) not null,
   LastName varchar(50) not null,
   OrderCount int not null default(0),
   FirstOrderDate datetime null,
   LastOrderDate datetime null
);
```

Sample records

CustomerID	FirstName	LastName	OrderCount	FirstOrderDate	LastOrderDate
20025	Jerome	Hatfield	1	2011-09-02 00:00:00.000	2011-09-02 00:00:00.000
20026	Kevin	Garza	0	NULL	NULL

```
dbcc traceon (3604);
dbcc page ('CorpDB', 1, 6406, 3);
```

DBCC PAGE output

```
Slot 123 Offset 0x1833 Length 51

Record Type = PRIMARY_RECORD        Record Attributes =  NULL_BITMAP VARIABLE_COLUMNS
Record Size = 51
Memory Dump @0x00000100FAF0B833

0000000000000000:   30001c00 394e0000 01000000 00000000 529f0000  0...9N..........R..
0000000000000014:   00000000 529f0000 06000002 002b0033 004a6572  ....R.......+.3.Jer
0000000000000028:   6f6d6548 61746669 656c64                      omeHatfield

Slot 123 Column 1 Offset 0x4 Length 4 Length (physical) 4

CustomerID = 20025

Slot 123 Column 2 Offset 0x25 Length 6 Length (physical) 6

FirstName = Jerome

Slot 123 Column 3 Offset 0x2b Length 8 Length (physical) 8

LastName = Hatfield

Slot 123 Column 4 Offset 0x8 Length 4 Length (physical) 4

OrderCount = 1

Slot 123 Column 5 Offset 0xc Length 8 Length (physical) 8

FirstOrderDate = 2011-09-02 00:00:00.000

Slot 123 Column 6 Offset 0x14 Length 8 Length (physical) 8

LastOrderDate = 2011-09-02 00:00:00.000
```

DBCC PAGE output

CustomerID	FirstName	LastName	OrderCount	FirstOrderDate	LastOrderDate
20025	Jerome	Hatfield	1	2011-09-02 00:00:00.000	2011-09-02 00:00:00.000

```
Record Type = PRIMARY_RECORD        Record Attributes =  NULL_BITMAP VARIABLE_COLUMNS
Record Size = 51
Memory Dump @0x00000100FAF0B833

0000000000000000:   30001c00 394e0000 01000000 00000000 529f0000  0...9N..........R..
0000000000000014:   00000000 529f0000 06000002 002b0033 004a6572  ....R.......+.3.Jer
0000000000000028:   6f6d6548 61746669 656c64                      omeHatfield
```

Status bits

CustomerID	FirstName	LastName	OrderCount	FirstOrderDate	LastOrderDate
20025	Jerome	Hatfield	1	2011-09-02 00:00:00.000	2011-09-02 00:00:00.000

```
Record Type = PRIMARY_RECORD        Record Attributes =  NULL_BITMAP VARIABLE_COLUMNS
Record Size = 51
Memory Dump @0x00000100FAF0B833

0000000000000000:   3000 1c00 394e0000 01000000 00000000 529f0000  0...9N..........R..
0000000000000014:   00000000 529f0000 06000002 002b0033 004a6572  ....R.......+.3.Jer
0000000000000028:   6f6d6548 61746669 656c64                       omeHatfield
```

Status Bits = 0x0030 = 0000 0000 0011 0000

Has NULL bitmap, has variable-length columns

Fixed-length data size

CustomerID	FirstName	LastName	OrderCount	FirstOrderDate	LastOrderDate
20025	Jerome	Hatfield	1	2011-09-02 00:00:00.000	2011-09-02 00:00:00.000

```
Record Type = PRIMARY_RECORD        Record Attributes =  NULL_BITMAP VARIABLE_COLUMNS
Record Size = 51
Memory Dump @0x00000100FAF0B833

0000000000000000:   3000 1c00 394e0000 01000000 00000000 529f0000  0...9N..........R..
0000000000000014:   00000000 529f0000 06000002 002b0033 004a6572  ....R.......+.3.Jer
0000000000000028:   6f6d6548 61746669 656c64                       omeHatfield
```

0x001c (28 bytes)

Fixed-length data

CustomerID	FirstName	LastName	OrderCount	FirstOrderDate	LastOrderDate
20025	Jerome	Hatfield	1	2011-09-02 00:00:00.000	2011-09-02 00:00:00.000

```
Record Type = PRIMARY_RECORD        Record Attributes =  NULL_BITMAP VARIABLE_COLUMNS
Record Size = 51
Memory Dump @0x00000100FAF0B833

0000000000000000:   30001c00 394e0000 01000000 00000000 529f0000  0...9N..........R..
0000000000000014:   00000000 529f0000 06000002 002b0033 004a6572  ....R.......+.3.Jer
0000000000000028:   6f6d6548 61746669 656c64                      omeHatfield
```

0x00004e39 = 20025

0x9f52 = 40786 (days past 1900-01-01)

Number of columns

CustomerID	FirstName	LastName	OrderCount	FirstOrderDate	LastOrderDate
20025	Jerome	Hatfield	1	2011-09-02 00:00:00.000	2011-09-02 00:00:00.000

```
Record Type = PRIMARY_RECORD        Record Attributes =  NULL_BITMAP VARIABLE_COLUMNS
Record Size = 51
Memory Dump @0x00000100FAF0B833

0000000000000000:   30001c00 394e0000 01000000 00000000 529f0000  0...9N..........R..
0000000000000014:   00000000 529f0000 0600[illegible]002 002b0033 004a6572  ....R.......+.3.Jer
0000000000000028:   6f6d6548 61746669 656c64                      omeHatfield
```

6 columns

NULL bitmap

CustomerID	FirstName	LastName	OrderCount	FirstOrderDate	LastOrderDate
20025	Jerome	Hatfield	1	2011-09-02 00:00:00.000	2011-09-02 00:00:00.000

```
Record Type = PRIMARY_RECORD        Record Attributes =  NULL_BITMAP VARIABLE_COLUMNS
Record Size = 51
Memory Dump @0x00000100FAF0B833

0000000000000000:   30001c00 394e0000 01000000 00000000 529f0000  0...9N..........R..
0000000000000014:   00000000 529f0000 06000002 002b0033 004a6572  ....R.......+.3.Jer
0000000000000028:   6f6d6548 61746669 656c64                      omeHatfield
```

No NULL data

Number of variable-length columns

CustomerID	FirstName	LastName	OrderCount	FirstOrderDate	LastOrderDate
20025	Jerome	Hatfield	1	2011-09-02 00:00:00.000	2011-09-02 00:00:00.000

```
Record Type = PRIMARY_RECORD        Record Attributes =  NULL_BITMAP VARIABLE_COLUMNS
Record Size = 51
Memory Dump @0x00000100FAF0B833

0000000000000000:   30001c00 394e0000 01000000 00000000 529f0000  0...9N..........R..
0000000000000014:   00000000 529f0000 06000200 2b0033 004a6572  ....R.......+.3.Jer
0000000000000028:   6f6d6548 61746669 656c64                    omeHatfield
```

2 variable-length columns

Variable-length columns offsets

CustomerID	FirstName	LastName	OrderCount	FirstOrderDate	LastOrderDate
20025	Jerome	Hatfield	1	2011-09-02 00:00:00.000	2011-09-02 00:00:00.000

```
Record Type = PRIMARY_RECORD        Record Attributes =  NULL_BITMAP VARIABLE_COLUMNS
Record Size = 51
Memory Dump @0x00000100FAF0B833

0000000000000000:   30001c00 394e0000 01000000 00000000 529f0000  0...9N..........R..
0000000000000014:   00000000 529f0000 06000002 002b0033 004a6572  ....R.......+.3.Jer
0000000000000028:   6f6d6548 61746669 656c64                      omeHatfield
```

Marks where each variable-length column data ends

Variable-length data

CustomerID	FirstName	LastName	OrderCount	FirstOrderDate	LastOrderDate
20025	Jerome	Hatfield	1	2011-09-02 00:00:00.000	2011-09-02 00:00:00.000

```
Record Type = PRIMARY_RECORD        Record Attributes =  NULL_BITMAP VARIABLE_COLUMNS
Record Size = 51
Memory Dump @0x00000100FAF0B833

0000000000000000:   30001c00 394e0000 01000000 00000000 529f0000  0...9N..........R..
0000000000000014:   00000000 529f0000 06000002 002b0033 004a6572  ....R.......+.3.Jer
0000000000000028:   6f6d6548 61746669 656c64                      omeHatfield
```

String data (either one or two bytes per character)

DBCC PAGE output

CustomerID	FirstName	LastName	OrderCount	FirstOrderDate	LastOrderDate
20026	Kevin	Garza	0	NULL	NULL

```
Slot 124 Offset 0x1866 Length 47

Record Type = PRIMARY_RECORD        Record Attributes =  NULL_BITMAP VARIABLE_COLUMNS
Record Size = 47
Memory Dump @0x00000100FAF0B866

0000000000000000:   30001c00 3a4e0000 00000000 00000000 02000000  0...:N..............
0000000000000014:   00010000 88c110fb 06003002 002a002f 004b6576  .....Á.û..0..*./.Kev
0000000000000028:   696e4761 727a61                               inGarza
```

NULL bitmap

CustomerID	FirstName	LastName	OrderCount	FirstOrderDate	LastOrderDate
20026	Kevin	Garza	0	NULL	NULL

```
Slot 124 Offset 0x1866 Length 47

Record Type = PRIMARY_RECORD        Record Attributes =  NULL_BITMAP VARIABLE_COLUMNS
Record Size = 47
Memory Dump @0x00000100FAF0B866

000000000000000:   30001c00 3a4e0000 00000000 00000000 02000000  0...:N..............
000000000000014:   00010000 88c110fb 06003002 002a002f 004b6576  .....Á.û..0..*./.Kev
000000000000028:   696e4761 727a61                               inGarza
```

NULL bitmap = 0011 0000

Ack! I can't read DBCC PAGE!

- Some feature will made DBCC PAGE not human readable
  - Transparent data encryption
  - Compression
  - Columnstore
  - In-Memory OLTP (good luck finding a page number)

Checkpoint

- Process of writing dirty pages from the buffer pool to disk
  - Irrespective of transaction completion

Checkpoint Types

- Automatic
  - Period background thread
  - Instance-wide [`sp_configure 'recovery interval (min)', 2`]
- Indirect (2012+)
  - Database-specific
  - [`alter database myDB set target_recovery_time = 2 minutes`]
  - Off by default in 2012, 2014; on by default in 2016+
- Internal
  - During operations such as backup, snapshots, shutdown
- Manual
  - CHECKPOINT command

Checkpoint Process

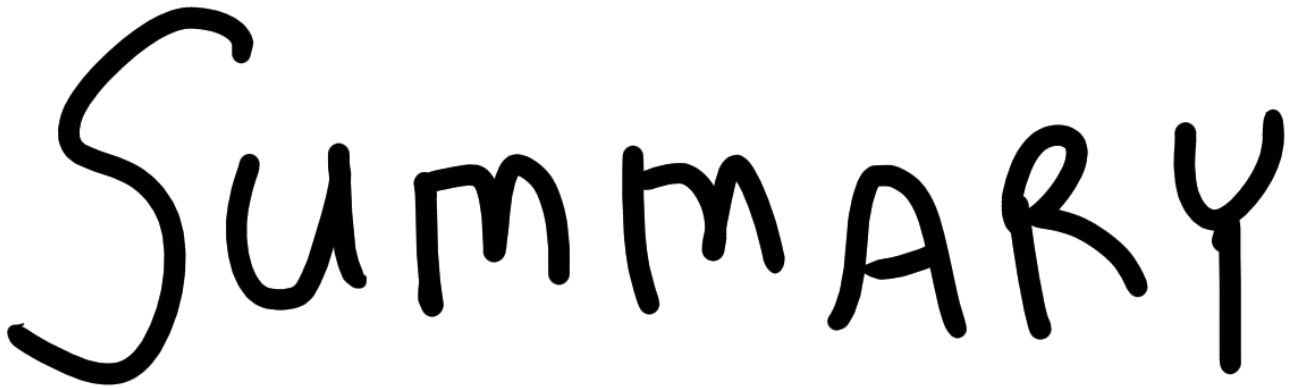
- Write to log: checkpoint start
  - Also info about any uncommitted transactions
  - Flush the log
- Identify dirty pages; write to disk
- Update boot page with LSN corresponding to checkpoint start
- (If SIMPLE recovery) clear the log
- Write to log: checkpoint finish

Part 7

Index pages

- B-Tree structure same as clustered index
  - Only key values in root and internal nodes
  - Included column data only in leaf nodes
- Always includes reference back to table
  - If table is a clustered index, includes the clustering keys (no duplicates!); may require “uniquifier”
  - If table is a heap, includes a “row identifier” (file:page:slot)

Epilogue



Summary

- Data files are organized into extents and pages
- Many page types
  - Several bitmaps to store allocation data
  - Miscellaneous pages (boot, file header, PFS)
  - Data pages and index pages

Summary

- Understanding storage internals will help
  - Better table design
  - More efficient use of storage systems
  - More efficient SQL operations (i.e., faster!)

Appendix

Commands

Command	Description	Example
DBCC PAGE *	Outputs contents of a page	`dbcc page ('CorpDB', 1, 2, 3);`
DBCC IND	Outputs pages associated to an index	`dbcc ind ('CorpDB', 'Customer', 1);`
DBCC FILEHEADER *	Outputs contents of file header page	`dbcc fileheader ('CorpDB', 1);`
DBCC DBINFO *	Outputs contents of boot page	`dbcc dbinfo;`

* Turn on trace flag 3604

Commands

Command	Description	Example
%%physloc%%	Virtual column indicating location of a row	`select *, %%physloc%% from table`
sys.fn_PhysLocFormatter	Formats %%physloc%%	`select *, sys.fn_PhysLocFormatter(%%physloc%%) from table`

Resources

- Paul Randal's "Inside the Storage Engine" series
  - Anatomy of a record - a page - an extent
  - IAM pages – Bitmap pages – Header pages – Boot page
- Kalen Delaney, *SQL Server 2008 Internals*

Thank You

- This presentation and supporting materials can be found at www.tf3604.com/internals.
  - Slide deck
  - Scripts
  - Sample database

brian@tf3604.com • @tf3604