

CS145: Study Guide

Abstract

These study notes are to make sure you understand the high-level concepts.

- **Please consult the slides first** as we've tried to make them a complete reference for the course.
- If you are confused, please ask questions on Piazza—we're happy to update these notes! (You'll notice below that there is a "Notes Compiled" time to keep track of versions.)
- These notes may contain more color about topics not discussed in lecture. Unlike lecture and homework, material covered in these notes is **NOT** necessarily required knowledge. However material that is not required knowledge will mostly be displayed in boxes which say "*Extra Material*"
- Finally, these are *VERY* rough notes, but we hope they help you!

Notes Compiled: Saturday 14th November, 2015 at 14:25

Contents

1	Transactions	4
1.1	The ACID Properties	4
1.2	TXNs In SQL	5
1.3	Logging for Atomicity & Durability	5
1.3.1	The Log	5
1.3.2	Write-Ahead Logging	5
1.4	Concurrency with Isolation & Consistency	6
1.4.1	Scheduling Definitions	6
1.4.2	Anomalies	6
1.4.3	Conflicts	6
1.4.4	Locking	6
1.4.5	Two-Phase Locking (2PL) Protocol	7
1.4.6	Conflict Equivalence	7
1.4.7	Dependency Graphs	7
1.4.8	Deadlock	8
1.5	Some Highlights and Tips	9
2	Storage	11
2.1	Disk Mechanics	11
2.2	Disk Latency	11
2.3	High-level: Disk vs. Main Memory	12
2.4	Cost Model	12
3	Indexing	13
3.1	Index Operations	13
3.2	Index Classifications	14
3.3	B+ Trees	14
3.4	The Height of the Tree	15
3.5	Setting d	16
3.5.1	Search	16
4	External Merge Sort	18
5	Join Algorithms	21
5.0.2	Notation	21
5.1	Block Nested Loop Join (BNLJ)	21
5.2	Index Nested Loop Join	23
5.3	Sort Merge Join	24

5.3.1	Optimization: Join and Merge!	25
5.4	Hash Join	26
5.4.1	Partitioning Phase	26
5.4.2	Join Phase	30
5.5	Join Algorithm Comparison	32
6	Histograms	32

1 Transactions

[Lectures 8-9]

A *transaction* (*TXN*) is a sequence of one or more operations (reads or writes), which reflect a single real-world transition (which as in the real world, should happen either completely or not at all). Grouping changes to the state of a database into TXNs allows:

- One or more users to run multiple queries concurrently, with a high level of performance, without having to take concurrency into account
- DBMS recovery from crashes or user-initiated aborts to a consistent state without any additional user supplied code.

1.1 The ACID Properties

- **Atomicity:** The DBMS state either reflects all effects of a transaction, or none of them.
- **Consistency:** If (explicit and implicit) integrity constraints hold over the state before a transaction, they will also hold over the state after the transaction
- **Isolation:** The effect of transactions concurrently is the same as the effect of running all transactions one-at-a-time
- **Durability:** Once a transaction has committed, its effects remain in the database (are stably stored on disk, flash, or other non-volatile storage)

ACID Challenges

- Power failures
- User-forced aborts/rollbacks
- Concurrency (DBMS can freely interleave individual actions of separate transactions)
- Performance

28 1.2 TXNs In SQL

- 29 • Defaults to one transaction per SQL statement
- 30 • START TRANSACTION to begin and COMMIT or ROLLBACK to
- 31 end can manually specify a transaction

32 1.3 Logging for Atomicity & Durability

33 1.3.1 The Log

34 **Goal:** Ensure that operations (e.g. from partially-completed TXNs) can be
35 undone, so that we can write partial TXNs to disk and still ensure atomicity

- 36 • A list of modifications: the basic idea is to record UNDO information
37 for every update, writing a diff (minimal info) to the log sequentially
- 38 • Log records contain (XID, location, old data, new data), which is
39 enough to undo any transaction (Here XID is the TXN id)
- 40 • Duplexed and archived on stable storage (we assume the log is not
41 lost)
- 42 • Can *force writes* to disk.
- 43 • Handled internally by the DBMS (transparent to users)

44 1.3.2 Write-Ahead Logging

- 45 • Basic protocol:
 - 46 1. All log records for the TXN force-written to disk
 - 47 2. Commit record written to disk & **TXN commits**
 - 48 3. Data written to disk
- 49 • Must force log record to disk for an update before the corresponding
50 data page goes to storage (guarantees Atomicity)
- 51 • Must write all log records for a transaction before commit. (guarantees
52 Durability)
- 53 • The commit is guaranteed when the commit record hits disk.

54 1.4 Concurrency with Isolation & Consistency

55 1.4.1 Scheduling Definitions

- 56 • A *schedule* is a particular interleaving of the operations of multiple
57 TXNs
- 58 • *Serial Schedule*: A schedule that does not interleave the actions of
59 different transactions
- 60 • *Equivalent Schedules*: Two schedules such that executing them have an
61 identical effect on the database state, for any possible initial database
62 state.
- 63 • *Serializable Schedule*: A schedule that is equivalent to *some* serial
64 schedule for the transactions scheduled.

65 1.4.2 Anomalies

- 66 • Lost update/Overwriting uncommitted data (Write after Write)
- 67 • Dirty read (Read after Write; writing transaction subsequently aborts)
- 68 • Inconsistent read (A task sees some but not all changes made by an-
69 other)

70 1.4.3 Conflicts

- 71 • Two actions *conflict* if they are:
 - 72 – Part of different TXNs
 - 73 – Involve the same variable
 - 74 – At least one is a write
- 75 • Note that a conflict is **not** an inherently bad occurrence, nor does it
76 imply an anomaly- just a definition

77 1.4.4 Locking

78 Ensure that each transaction's view of the DB state is 'consistent.'

79 1.4.5 Two-Phase Locking (2PL) Protocol

- 80 • In 2PL, there are two phases of locking: lock acquisition and lock
81 release. Once TXN releases a lock, it cannot acquire any more locks.
- 82 • Each transaction must obtain a *S (shared) lock* on an object before
83 reading it, and a *X (exclusive) lock* on an object before writing it
 - 84 – An S lock can be held by multiple TXNs
 - 85 – An X lock can be held by only one TXN
 - 86 – An X lock cannot be held on an object at the same time as any
87 other S or X locks on that object
- 88 • All locks held by a transaction are released when the transaction com-
89 pletes.
- 90 • We use Strict 2PL in this course, which has the added requirement over
91 2PL that a transaction releases locks only at the end of the transaction.

92 Strict 2PL guarantees that schedules are conflict serializable, but does
93 not allow all serializable schedules. See lecture for diagrams comparing these
94 two protocols.

95 1.4.6 Conflict Equivalence

96 *We define conflict equivalence in order to capture the behavior of locking,*
97 *keep that in mind and this may make more sense!*

98 Two schedules are *conflict equivalent* if:

- 99 • The schedules involve the same actions of the same transactions
- 100 • Every pair of conflicting actions of the two TXNs are ordered in the
101 same way

102 A schedule is *conflict serializable* if it is conflict equivalent to *some* serial
103 schedule. Conflict serializable schedules are a subset of serializable schedules.

104 1.4.7 Dependency Graphs

105 A dependency graph consists of a directed graph with:

- 106 • One node per transaction $T_1 \dots T_N$

- An edge from T_i to T_j if they conflict on an object and the action of T_i precedes T_j .

With Strict 2PL, a schedule is *conflict serializable* if and only if its dependency graph is acyclic.¹ To see the forward direction, observe that a conflict serializable graph that uses Strict 2PL and does not deadlock ensures that there cannot be a cycle in the dependency graph: All locks are held until the end of the transaction. Hence, if there were a cycle this would cause a deadlock. On the other hand, if the dependence graph for a schedule S is acyclic, then one can use any topological order for the graph that agrees with the edges to create a serial schedule. Notice that S is conflict equivalent to the serial schedule, since each conflicting action is ordered in the same way. Hence, S is conflict equivalent to a serializable schedule, hence conflict serializable.

One possibly helpful way to remember this is: *strict 2PL (and locking in general) converts bad (cyclic) conflicts into deadlocks*. We describe deadlocks more next.

1.4.8 Deadlock

A **deadlock** is a cycle of transactions waiting for locks to be released by one another (and thus unable to proceed).

Example 1.1 *Here is a schedule that could deadlock:*

$$T_1 : W(A); \quad T_2 : W(B); \quad T_2 : R(A); \quad T_1 : R(B)$$

If we use locking, the deadlock occurs at the last transaction. Here, T_1 holds the lock on A , which causes T_2 to block waiting for T_1 . Before blocking, T_2 has acquired an exclusive lock on B . T_1 attempts to acquire this lock from T_2 , but now it blocks. Neither can make progress, and they are deadlocked.

¹This is true more generally, but we did not consider those issues in this course.

Extra Material: Deadlock Prevention and Detection

There are two common ways to handle deadlock: Deadlock Prevention and Deadlock Detection. Detection is more common.

Deadlock Prevention Assign priorities to transactions based on timestamps. Lower timestamps have higher priority.

- If transaction T_i wants a lock held by T_j , then act on either a *wait-die* or *wound-wait* policy.
- Wait-die: If T_i has higher priority, T_i waits for T_j ; otherwise T_i aborts.
- Wound-wait: If T_i has higher priority, T_j aborts; otherwise T_i waits.
- In either case, make sure that restarted transactions keep their old timestamp, so they eventually have the priority to succeed.

Deadlock Detection We create a *waits-for graph* and check this graph for cycles periodically.

- There is a single node for each transaction. We will abuse notation and conflate the node and the transaction it represents.
- There is an edge from T_i to T_j if T_i is waiting for T_j to release a lock.
- If the graph has a cycle, then there may be deadlock²
- Periodically check for cycles. If a cycle is found, kill a transactions on the cycle to break the deadlock.
- The check for a cycle and deadlock could be expensive if there are large cycles. However, typically the cycles in the graph tend to be small, and one can optimize for this.

131

132 1.5 Some Highlights and Tips

133 Note there are examples in the Lectures 8-9. Check them out! Serializability
134 is the database gold standard notion of correctness. It allows us to inter-

135 leave the order of transactions for performance. Conflict serializability is a
136 restricted notion that captures what our locking protocol allows.

- 137 • Intuitively, to see if a schedule is equivalent to a fixed serial schedule
138 transactions- say T_1, \dots, T_k - you simulate running these transactions.
139 You check that for each object they write, the schedule and the serial
140 schedule have the *same value at the end of both schedules*. It does
141 not matter if the two databases have different values in the middle of
142 execution!
- 143 • To test if a schedule is *serializable*, you can do the following. Sup-
144 pose the schedule contains k transactions, T_1, \dots, T_k . You consider
145 whether that schedule is equivalent to *some* ordering of the transac-
146 tions. You simply try the algorithm from the preceding bullet point on
147 each possible orders (all $k!$ orders). This is definitely a tougher notion
148 to test!
- 149 • If no such ordering exists, the transaction is *not* serializable. Intu-
150 itively, it means that there is some kind of dependence like in Lecture
151 12, where some examples illustrate this.
- 152 • To test if a schedule is *conflict serializable*, you simply run the strict
153 2PL locking protocol and see if you get stuck! More precisely, you
154 form the conflict graph and check if it is acyclic.
- 155 • Not all serializable schedules are conflict serializable. This makes sense
156 as conflict Serializability is a more strict notion than serializability!
- 157 • You could still have a deadlock even when running Strict 2PL. If you
158 don't have a deadlock, and you run strict 2PL to completion for a
159 schedule, then your schedule is serializable. This captures the sense in
160 which locking is correct.

161 2 Storage

162 [Lecture 12]

163 We study hard disks, which we assume are durable (we do not worry
164 about failure, which is handled by other mechanisms). Disks are block
165 devices.³ This means we can only read or write data at block sizes (4K,
166 8K or so). We need to bring the data into memory to modify it or have
167 the CPU read it, a region of memory called the buffer. The buffer may be
168 smaller than all available memory.

169 2.1 Disk Mechanics

170 **These are not important for the exam.**

- 171 • Disk storage hardware is comprised of multiple *platters*.
- 172 • A *track* is a ring in a platter (bytes per track on the order of 10^5)
- 173 • A *cylinder* is one track from every platter (the track in the same
174 location on each platter)
- 175 • Read and write from the disk in units of disk blocks (typically 4K, 8K,
176 16K)

177 2.2 Disk Latency

178 **These are not important for the exam.**

- 179 • Disk Latency: Total time from when command is issued to when data
180 is in memory
- 181 • Comprised of *seek time* + *rotational latency*
- 182 • Sequential disk access much more efficient than random disk access
- 183 • Seek time comes from the time needed for the disk head to reach the
184 track on which the block to be read lies

³We will conflate blocks and pages for this course.

185 2.3 High-level: Disk vs. Main Memory

- 186 • Accessing (reading from / writing to) the disk is slow, whereas access-
187 ing main memory is fast; in particular:
 - 188 – Main memory is $\approx 10x$ faster for *sequential access*
 - 189 – Main memory is $\approx 100,000x$ faster for *random access* (i.e. not
190 sequential)
- 191 • Disk is *durable* (we assume data is safe once on disk for this course)
192 while main memory is *volatile*- data can be lost if crash occurs, etc.
- 193 • Main memory is far more expensive per unit of storage space than
194 disk- we'll assume in this course that we have (effectively) unlimited
195 disk space but limited main memory

196 2.4 Cost Model

197 We will count cost in what is called the *IO cost model* throughout the rest
198 of these notes: We count the number of pages read and written to disk and
199 ignore CPU costs. These costs are a simplification of the true costs sufficient
200 to capture the rough trends. The message of these costs models is that it is
201 indeed possible for the machine to estimate your cost; there are some hard
202 elements like statistics, but we exposed you to them so you would know how
203 to write efficient queries.

204 **Remark 2.1** *The buffer is not all the memory of the machine (there are*
205 *registers, a small amount of scratch space, etc.). It is however the place*
206 *where all data pages are accessed. We count only access to the data pages*
207 *themselves in our model. We will also not care too much about the mech-*
208 *anics of the buffer pool, but its size can be important for performance.*

209 3 Indexing

210 [Lectures 13-14]

211 An index on a file speeds up selections on the search key fields of the
212 index. An index is stored in a separate file (in pages). In particular, there
213 is a data file and there is an index that points into that data file.

214 **Search Key** The search key fields:

- 215 • Can be any subset of the relation's fields
- 216 • Is not the same thing as a key of the relation

217 Indexes support efficient retrieval of all data entries k^* with a particular
218 value k for the search key field.

219

Extra Material: Page IDs and Record IDs (rids)

220 A Page ID is a page identifier. A record id (rid for short) are in-
formally called pointers in these notes. *These are **not** exactly like
pointers in C.* An rid encodes the location of record unambiguously
on secondary storage (e.g., which disk in which file at which point
in that file.) In contrast to a C pointer, you can store a pageid or
an rid on disk, and it remains valid.

221 **Contents of an Index** There are three main options for what the index
222 stores for each k :

- 223 1. the actual record
- 224 2. (k, rid) : the key plus the record id (with duplicates, duplicates are
225 stored).
- 226 3. $(k, \text{rid list})$

227 Typically only one index uses choice 1, to avoid duplicating the actual
228 records. We will consider only the case of k and a single rid in this course.

229 3.1 Index Operations

230 The valid operations on an index are:

- 231 • Search: Given a key k find all records that match a key.

- 232 • FindRange: Given a pair of keys $[k_1, k_2]$ find records that have keys in
233 this range.
- 234 • Insert and Remove (we will not study these in detail)

235 3.2 Index Classifications

236 An index is *clustered* if the data is ordered in the same way as the underlying
237 data. Otherwise, we say it is unclustered. Whether the index is clustered
238 or not has an enormous impact on query performance, for *range queries*.
239 The key reason is that when moving from the leaf pages, to the data we will
240 potentially do random IOs and read many pages. This calculation is in the
241 lecture slides!

242 3.3 B+ Trees

- 243 • Very good for range queries and sorted data.
- 244 • These are search trees, but the B refers to neither ‘Binary’ nor ‘Bal-
245 anced’ (although B and B+ trees are balanced).
- 246 • For B+ trees, the basic idea is to have the leaves of the tree be a
247 linked list of the physical pages. This is to support efficient scanning
248 of a range of values.
- 249 • The parameter d of a B+ tree is called its order, which we describe
250 below.
- 251 • Both internal and leaf nodes are set to be the size of a single page. See
252 Lectures 13-14 for intuition as to why.

253 **Internal Nodes** See the slides for pictures of these structures.

- 254 • Each node has $\geq d$ and $\leq 2d$ search keys—except for possibly the root.
255 These search keys are also called *guard entries*.
- 256 • Each has a pointer (PageID) to that points to a node for each range
257 of key values that lie between the values of the keys in the node. In
258 particular, if we have two guard entries g_1 and g_2 with a pointer p
259 between them p contains search keys in the range $[g_1, g_2)$.

260 Leaf Nodes

- 261 • Each leaf also has between $\geq d$ and $\leq 2d$ search keys.
- 262 • Each search key is paired with a single rid.
- 263 • There is also a pageid that points to the next page next leaf node;
264 this way the leaf nodes form a linked list that can be more cheaply
265 traversed for range queries than descending from the root repeatedly.

266 3.4 The Height of the Tree

267 Let's give some bounds on how tall the tree is to exercise notation. Suppose
268 we have N leaf pages. We want to compute the height h .

- 269 • **Sparsest Tree (Take 1)** Suppose the suppose the root has at least
270 $d + 1$ pointers. Since we know that each internal node has at least $d + 1$
271 pointers, and Then, the height of the tree h must satisfy must satisfy
272 $(d + 1)^h \leq N$.
- 273 • **Sparsest Tree (Take 2)** If the root does not have $d + 1$, it must have
274 at least two pointers (else you could simply remove it!). In that case,
275 the minimum number of entries is $2(d + 1)^{h-1} \leq N$.
- 276 • **Densest Tree.** On the other hand, there are at most $2d + 1$ pointers
277 per internal node and $2d + 1$ entries per leaf node, hence $N \leq (2d + 1)^h$.
278 So we have that d , h , and N are in the following relationship.

$$2(d + 1)^{h-1} \leq N \leq (2d + 1)^h$$

279 **Fill Factor** To get a better sense of how tall a tree is in practice, we
280 observe that most nodes are filled between 66 – 80%; we call this percentage
281 the *fill factor*. This room is left to amortize the cost of inserting new data
282 into the tree. This allows us to have intuition about the height of the tree
283 in practice (see the lecture slides for more examples).

- 284 • If the leaves are filled with a factor f , this means N pages of records
285 takes N/f pages. So with $f = \frac{2}{3}$ (in general a decent estimate), we
286 need $1.5N$ pages to store these records.
- 287 • In fact, nodes at higher levels of the trees tend to be more densely
288 packed than the leaves. Hence, leaf nodes could be filled around 66%
289 (to allow more slack for insertions) while internal nodes are filled to a
290 greater extent.

- This information regarding fill factor specifically is to help with intuition and will not be on the final exam. (The height is fair game).

3.5 Setting d

- d must be small enough that a B+ tree node can fit in one page in memory. On the other hand, we want the node to contain as many nodes as possible to obtain higher fanout.
- All but the last level of tree typically stay in the buffer pool, *which we do not account for in this class*. In some common cases, only I/O is actually needed to fetch a single record. The fanout is $2d + 1$ times the fill factor. If the fanout is F , then you have F times more data in the leaves than in the internal nodes. Often, $F \geq 100$, we have orders of magnitude differences in how much data we index on disk versus in memory.
 - A common example of this is the case in which we store all the data (not just the rid) in a leaf node. We won't consider this further.
 - A second example is when all the fields we want are part of the search key; we say the index is covering. E.g., we have an index with search key (A, B, C) and we ask a query like `SELECT B FROM A where A = 10`.
- This is covered explicitly in the lecture slides. However, the order is a function of the size of the key (and the size of a pageid).
- If we store the key and a pageid, the number of entries in a leaf page is a function the size of the key and the size of a page id (to point to the data page).

3.5.1 Search

To search a B+ tree:

- For Search: start at the root and descend down following pointers to the correct leaf. This is illustrated in the slides. As described above, when search for k we go down the first pointer such that $k \in [g_1, g_2)$.
- For a range query $[k_1, k_2]$, we search for k_1 as above. Then we scan along the linked list of leaves for the rest of the range. *This is where we use the pointers described in lecture.*

- The animations explain illustrate how this procedure works see lecture 13. (Animations require powerpoint or video.)

326 4 External Merge Sort

327 [Lectures 14-15]

328 The goal of external sort is to sort a file that is much larger than available
329 memory. This is *the standard merge sort algorithm* adapted to data on disk.
330 In particular, we will count IOs not CPU operations—since this captures the
331 dominant cost in dealing with IO. The algorithms are cheaper according to
332 this metric than say quicksort, which is preferred in an in-memory setting.

333 **Using More Memory** In lecture, we showed how to sort an arbitrarily
334 large file with only three buffer pages.⁴ We then asked the question: *If we*
335 *have more than three buffer pages, say $B+1$, how do we use these additional*
336 *buffer pages to improve the performance of our algorithm?* We used these
337 buffer pages in two ways: (1) longer initial sorted data and (2) merging runs
338 B -way rather than 2 way.

339 **External Merge Sort** The input is a single file that contains N pages
340 and our goal is to output that file in sorted order. We also have $B+1$ buffer
341 pages of memory. For concreteness, we will think about sorting the records
342 in ascending order. The algorithm runs in two phases: (1) creating initial
343 runs, and (2) merging those runs. *We will use the merging phase later in*
344 *sort-merge join.*

345 • **Phase 1: Create initial runs.** In the first phase of the algorithm,
346 we sort as much data as we can fit into memory. In this case, we can
347 hold $B+1$ buffer pages. We read all N pages but in $B+1$ sized chunks;
348 we sort these chunks in memory (using quicksort, say); and then we
349 write each now-sorted chunk back to the file. We call each such file
350 a *run*, which is a list of records in sorted order. These runs partition
351 the input. The output of Phase 1 is several runs. In more detail,

352 – Each run is of size $B+1$, and
353 – $\lceil \frac{N}{B+1} \rceil$ runs created in the first step, since the runs form a parti-
354 tion.

355 • **Phase 2: Merge** The input to this phase is some number of sorted
356 runs, and we merge the sorted runs. We will produce runs that are

⁴Here, the sorting time was $2N(1 + \log_2 N)$, in the terminology below our initial runs are of length 1 and we are doing a 2-way merge.

longer/larger, and we will reduce the number of sorted runs on each pass. We will run several passes of merging.⁵ The main observation is

To merge k sorted lists, we only need to hold the first element of each list in memory—no matter how long the list is.

A run is just a list. To find the smallest element in k lists, we just need the smallest element from each list (the first element of the sorted list.) Thus, with $B + 1$ buffer pages, we can perform a B -way merge:

- We use B pages, one for each of the pages, which contains the smallest elements of each run that has not been merged.
- We use one page for output. When the output fills up, we write it to disk.
- The length of the output run produced by merging runs of length N_1, \dots, N_B is $\sum_{i=1}^B N_i$, i.e., the sum of the size of the input runs.

See Lecture 12 for more detail on the external merge algorithm which constitutes this merge phase.

Analyzing the number of passes Once we have a run of length N then we've sorted the file! Let's analyze how many passes of the above merging we need to complete this task. We can think about it in two different ways: one by analyzing the number of runs left after each merge step and one by the length of the runs.

• Method 1: Number of runs.

- Every merge pass, we reduce the number of runs by a factor of B (we replace them with longer runs!)
- Thus, to analyze external merge sort, we want to know when we have reduced the number of runs to ≤ 1 . Roughly this is when we have done m passes

$$B^{-m} \left\lceil \frac{N}{B+1} \right\rceil \leq 1 \text{ or } \log_B \left\lceil \frac{N}{B+1} \right\rceil \leq m$$

We pick the smallest value of m that satisfies this inequality, as our goal is to do this in the fewer number of passes.

- Thus, we need to do $\lceil \log_B \lceil \frac{N}{B+1} \rceil \rceil$ passes to fully merge.

⁵This is described at least twice in lecture with fancy animations, please look at it.

386 • Method 2: Length of Runs. We will assume for simplicity that all
 387 initial runs are the same size, $B + 1$.⁶

388 – If we merge B runs of length R , we get a new run of length BR .
 389 Thus, after m merge passes we could have runs as large as

$$B^m R$$

390 – We have to repeat this process until we have one run that is the
 391 length of the entire file. Thus, ignoring ceilings, after m merge
 392 passes we need to solve:

$$B^m(B + 1) \geq N$$

393 Thus, $m \geq \log_B \lceil \frac{N}{B+1} \rceil$ and again we need to do $\lceil \log_B \lceil \frac{N}{B+1} \rceil \rceil$
 394 passes.

395 **IO Cost** The key observation is that in both phases on each pass, *each*
 396 *page is read once and written exactly once for 2 IOs per page*. Thus, our
 397 number of IO operations is $2N(1 + m)$ where 1 is the initial pass and m is
 398 the number of merge passes. In other words our total IO cost is:

$$2N \left(1 + \lceil \log_B \left\lceil \frac{N}{B+1} \right\rceil \right)$$

399 **Further optimization: Repacking** In Lecture 13 we outline a further
 400 optimization where we "merge" in the buffer as we sort at the initial stage,
 401 to create longer initial runs. As a rough estimate, using this technique we
 402 can get initial runs of size $\approx 2(B + 1)$ and then our total IO cost is:

$$2N \left(1 + \lceil \log_B \left\lceil \frac{N}{s(B+1)} \right\rceil \right)$$

403 Note that since this improvement is approximate, we'll explicitly say if
 404 repacking is being used (unless we do, in any of the calculations below,
 405 assume it is not)

⁶If you do not make this assumption, the length of the merged file is the size of the sum of the input files. The details are not too nasty. In particular, for m , all but one run is of length $B^m(B + 1)$. That is, there is exactly one run that is smaller in any pass.

406 5 Join Algorithms

407 We described three join algorithms: block nested loop join (see slides), Hash
408 Join, and Sort Merge Join. We will only talk about equijoins here, so it
409 makes sense to talk about a join key. For example, we could use it on a
410 query like the following:

```
411 SELECT * FROM Employee e, Dept d where e.did = d.id
```

412 Or more abstractly.

```
413 SELECT * FROM R, S where R.A = S.A
```

414 We will use A as the join key throughout (although it could be a con-
415 junction of more than one equality atoms e.g., $R.A = S.A$ AND $R.B = S.B$).

416 5.0.2 Notation

417 We'll use the following notation as in lecture; for a relation R :

- 418 • $P(R)$ is the number of *pages* or *blocks* that make up R ⁷
- 419 • $T(R)$ is the number of tuples in R
- 420 • OUT represents the IO cost to write the output of the join back to
421 disk. In the worst case $OUT = \frac{T(R)T(S)}{P}$ where P is the number of
422 tuples that fit on an output page, however in general OUT is much
423 closer to $O(P(R) + P(S))$. Either way, **OUT will be the same for**
424 **any join algorithm**, so is not important to any comparisons we do

425 5.1 Block Nested Loop Join (BNLJ)

426 You should know the cost equations for nested loop join (NLJ) and block
427 nested loop join (BNLJ). It is implicitly used below. If you have $B+1$ buffer
428 pages, and you join relation R (having $P(R)$ pages) and relation S (having
429 $P(S)$ pages). This algorithm is described in Algorithm 1

430 **Memory** We allocate our buffers as follows:

- 431 • We read in chunks R in size $B - 1$.
- 432 • We read in one page of S one at a time using 1 buffer page.
- 433 • We have 1 page for output. As it fills up, we write it out to disk.

⁷Again, in this class we will conflate *page* and *block*, and use the former in this section

```

BNLJ(R : Relation, S : Relation, B : Integer);
Data: Relation R, a Relation S, and B + 1 buffer pages.
Result: Compute  $R \bowtie S$ 
Setup a single output page;
foreach Block of B - 1 pages of R do % B blocks read at a time
    foreach page PS of S do % this is a read from disk
        foreach tuple s ∈ PS do % The part below is in memory
            foreach tuple r in those B - 1 pages do
                if r[A] = s[A] then Write (r, s) to the output page if
                    output page is full then Write output page to disk
                end
            end
        end
    end
end
if Output page is not empty then Write the output page to disk
Algorithm 1: Block Nested Loop Join

```

434 **IO Cost** The cost of this algorithm is:

$$\left\lceil P(R) + \frac{P(R)}{B-1} \right\rceil P(S) + \text{OUT}$$

435 Notice that this formula is not symmetric. You should be aware of how to
436 minimize it.

- 437 • Notice that if $P(R) < B - 1$ then BNLJ has IO cost $P(R) + P(S) +$
438 OUT, i.e. it's linear
- 439 • BNLJ can be used with essentially arbitrary join predicates, e.g., in-
440 equality join conditions such as $R.A > S.B + 5$ or $f(R.A, S.B)$ where f
441 is an arbitrary boolean function. So it's very useful. In contrast, hash
442 and sort-merge join are for equijoins (joining on equality constraint(s)
443 only) and natural joins (joining on equality of all shared attributes).
444 On the one hand this limits what joins what we can do, however we
445 can be much smarter about these ones!
- 446 • Equijoins are more common. For the rest of this section, we can think
447 of the query $R(A, B) \bowtie S(A, C)$.

448 5.2 Index Nested Loop Join

449 We use an index to avoid scanning the entire relation as shown in Algo-
 450 rithm 2. We assume that we have an index on S with search key that
 451 contains (at least) A . The algorithm is as follows:

```

foreach tuple of R do % Read R one page at a time
    | Search using  $r[A]$  into the index on  $S$ ;
    | Output all matches returned by the index;
end

```

Algorithm 2: Index Nested Loop Join

452 The cost of this algorithm is as follows: we read in each page of R exactly
 453 once, so this incurs a total cost of $P(R)$. For each such tuple, we perform
 454 one index lookup in S . Thus, our cost can be written:

$$P(R) + T(R)\text{IndexLookup}(S) + \text{OUT}$$

455 Some comments on the algorithm:

- 456 • In our simplified model, the $\text{IndexLookup}(S)$ is proportional to the
 457 height of the tree, and so the number of leaf pages in S . *This level of*
 458 *detail is all you are required to know.*
- 459 • This algorithm requires three buffer pages to run: One for the input,
 460 one the output, and one for the index page. However, in practice, we
 461 can use the extra buffer pages to facilitate obvious caching opportuni-
 462 ties. For example, e.g., if the same value of $r[A]$ is queried twice, we
 463 could have cached all pages to the leaf in the index, which could fit in
 464 the buffer.

465 5.3 Sort Merge Join

466 The simple algorithm is roughly as follows: *sort using the external merge*
467 *sort, intersect the resulting lists, and output all matches.*

468 The sort phase is just external merge sort run on each of the relations.

469 Now, we have two relations R and S sorted by the join key A . For a
470 tuple $r \in R$, let $r[A]$ denote the value of the join key. Similarly, let $s[A]$
471 denote the value of the join key for a tuple $s \in S$.

472 **Merge For Join** We run essentially the same merge algorithm but now
473 over two files. Our goal is to find matching keys (and discard non matching
474 tuples). In essence, we are computing a list intersection with sorted lists.
475 All comparisons are with respect to the A value. That is,

- 476 • Find the minimum element of R in A -order, call it r .
- 477 • Find the minimum element of S in A -order, call it s .
- 478 • If $r[A] < s[A]$, then we know that no tuple remaining in S can possibly
479 match with r . In this case, we advance to the next tuple in r . (If
480 $r[A] > s[A]$, then the symmetric statements hold.)
- 481 • If $r[A] = s[A]$, then we have found a match. We write it to the output
482 buffer.
 - 483 – If there are multiple tuples with the same join key in R , S or
484 both, then we need to output the cross product of all of these
485 tuples. We find all tuples in R and S with join key equal $r[A]$
486 (and so also s).
 - 487 – Notice this cross product could be a huge number of tuples. In the
488 worst case, *we could even get an entire cross product as output.*
- 489 • We then advance through each list.
- 490 • **Backup:** Note that if there are many duplicate join keys, we may
491 have to "back up" and read in tuples we've already read in again. See
492 the animation in Lecture 15.

493 This is just the merge algorithm with a small modification. We are not de-
494 scribing any of the bookkeeping for merging a list, but it should be straight-
495 forward how it works (and you have animations from lecture). To do this
496 merge, we only need three pages. We use one page of R and one page of S
497 and use the algorithm from lecture.

498 **The Cost** With this simple algorithm to join R and S our cost is:

$$\underbrace{2P(R) \left(1 + \lceil \log_B \left\lceil \frac{P(R)}{B+1} \right\rceil \right) + 2P(S) \left(1 + \lceil \log_B \left\lceil \frac{P(S)}{B+1} \right\rceil \right)}_{\text{sorting cost}} + \underbrace{P(R) + P(S)}_{\text{merging cost}} + \text{OUT}$$

499 Note here we're explicitly writing OUT. How large can it be? Note also
500 that due to backup (mentioned above), the merging cost could actually be
501 larger (but usually won't be).

502 5.3.1 Optimization: Join and Merge!

503 We can do a simple optimization to take advantage of having more memory
504 in the final merge phase. In some cases, this may shave off an entire pass
505 over the data. If the buffer is large enough, then we can skip the last merge
506 step for the sorts of R and S individually. Instead, we can do their final
507 merge step and the join at the same time. In particular, with $B + 1$ buffer
508 pages, if the sum of the number of runs from both R and S is less than
509 than B ,⁸ we simply run the merge step for each in parallel. In particular,
510 we bring in one page from each of the runs in both R and S , and perform
511 the merge to find the smallest tuple (and run the merging step of the join
512 algorithm above). This shaves off the pass we would spend writing out the
513 fully sorted versions of R and S , only to merge the two lists again.

514 So under what conditions does this stroke of good luck occur to let us
515 sort in two passes? Suppose that we have $B + 1$ buffer pages. We just plug
516 in the number of runs we would have using the standard sorting algorithm.
517 The number of runs satisfy the above condition when:

$$\left\lceil \frac{P(R)}{B+1} \right\rceil + \left\lceil \frac{P(S)}{B+1} \right\rceil \leq B$$

518 In particular, very roughly, as long as $P(R) + P(S) \leq B^2$ we can do the
519 entire sort merge join in two passes. We can do slightly better below to
520 recover the statement that $\max\{P(R), P(S)\} \leq B^2$.

521 **Further Optimization** You can see that we can play similar games to
522 save one pass over R or one pass over S on the final merge run. It makes
523 the cost formula nastier, although it should be easy to see that you could
524 write code to estimate this formula.

⁸Note we need one page for output as is standard.

525 **Repack: Tournament Sort for Longer Runs** It turns out one can
 526 create initial sorted runs of length $\approx 2(B+1)$ using *tournament sort*.⁹ From
 527 the above inequality, this means that you finish in two passes when

$$\frac{P(R) + P(S)}{2} \leq \max\{P(R), P(S)\} \leq B^2$$

528 This factor of 2 is helpful to make the comparison to hashing easier.

529 5.4 Hash Join

530 The first phase of the Hash Join (HJ) algorithm is to partition into small
 531 enough buckets that the buckets from a smaller relation can fit in memory.
 532 A bucket is defined operationally: *all the values in the bucket will have*
 533 *the same hash value for the join key A*. In particular, the buckets form a
 534 partition of the input file so that if two tuples have the same join key, they
 535 will go to the bucket.

536 5.4.1 Partitioning Phase

537 We first describe the partitioning phase in Algorithm 3.

538 **Setup** We assume we have an infinite family of hash functions h_1, h_2, \dots ,
 539 that are distinct.¹⁰ A **bucket** is a file that consists of a set of pages. Only
 540 one page from a bucket will be in memory at any one time in Algorithm 3.

541 **Partitioning Pass** Given an input relation R with $P(R)$ pages and a size
 542 bound T , produce buckets so that each bucket contains no more than T
 543 pages.¹¹ The main operation in the partitioning phase is given $B+1$ buffer
 544 pages partition the input data into B buckets.

- 545 • We hash each tuple $r \in R$ on $r[A]$. This implies that all tuples with
 546 the same value of A are in the same bucket.
- 547 • All values in a bucket have the same of value the hash; this is the
 548 definition of a hash bucket.

⁹http://en.wikipedia.org/wiki/Tournament_sort

¹⁰We can be precise about, but it's painful. The idea is that two hash functions will hash different tuples to different places.

¹¹This may not always be possible, if for example all values are the same. Let's ignore this very minor detail that may obscure understanding. We will assume that the number of distinct values fit on T pages.

```

Partition( $R, A, \bar{h}, T$ );
Data: Relation  $R$ , an attribute of  $R$  called  $A$ , a family of hash
        functions  $\bar{h}$ , and a bucket size  $T$ .
Result: A set of buckets. Each bucket contains no more than  $T$ 
        pages
Buckets  $\leftarrow \{R\}$ . % initially the file is a single bucket;
 $i \leftarrow 1$  % this is which hash function to use;
while There is a Bucket with more than  $T$  pages do
    foreach Bucket  $P$  in Buckets do
        Create  $B$  new buckets  $P_1$  to  $P_B$  % each one will have one page
        in buffer.;
        foreach tuple  $t \in P$  do % Read 1 page of bucket  $P$  at a time
            Let  $j = h_i(t[A])$  % Use the hash function for this round.;
            Add the tuple  $t$  to the buffer page for bucket  $P_j$ ;
            if If the buffer page for  $P_j$  is full then
                Write  $P_j$  to disk;
                Clear contents of memory  $P_j$  in memory
            end
        end
        Write out all non-empty buckets.;
        Add new partitions to the set of all partitions.
    end
     $i \leftarrow i + 1$  % Advance to the next round.
end

```

Algorithm 3: Hash Partition

- 549 • We use 1 buffer page for input and B buffer pages as output (one page
550 per bucket). As the output buckets fill up, we write them to disk. As
551 a result, a bucket may be of size up to $P(R)$ (if the data are highly
552 skewed).
- 553 • If we assume “no skew”, operationally this means that each bucket
554 will be of roughly equal size after hashing. We will assume there
555 is no skew. Hence, the number of pages in each bucket under this
556 idealized assumption is $\lceil \frac{P(R)}{B} \rceil$. We make this *uniform hashing or no*
557 *skew assumption below*.

558 We can repeat the partitioning above to further partition the data. In the
559 next phase, we take each input bucket one at a time and treat it as a new
560 input. That is, we partition the data using a new hash function h_2 . Each

561 bucket produces B new buckets, hence The output is a set of B^2 buckets. We
 562 can repeat this process several times. After m passes, this process creates
 563 B^m buckets.

- 564 • Observe that on each pass, we use a distinct hash function.¹² We use
 565 hash function h_i on round i . Note that two tuples s, t are in the same
 566 bucket after m passes if:

$$h_i(s[A]) = h_i(t[A]) \text{ for all } i = 1, \dots, m$$

567 Alternatively, for two tuples s and t if there is even one i such that
 568 $h_i(t[A]) \neq h_i(s[A])$ for $i \in \{1, \dots, m\}$ then, s and t hash to different
 569 buckets.

- 570 • We also hope the size of the bucket goes down: *here is where we use*
 571 *the distinct hash function assumption*. If we reused the same hash
 572 function, the size of a bucket would not decrease after the first round.
 573 If we hash with a different hash function h_2 , however, then this allows
 574 us to further split the bucket. As a result, the size of each bucket will
 575 again go down by a factor of B . We repeat this process, and we use a
 576 distinct hash function on each pass.

577 **How many passes?** Notice after each pass, we introduce a factor of B
 578 more buckets and the size of each bucket is reduced by a factor of B (look
 579 familiar?). Hence, if we start with $P(R)$ pages of data, after m passes the
 580 size of a bucket is: $B^{-m}P(R)$.

- 581 • To ensure that $B^{-m}P(R) \leq T$, we again need to find the smallest
 582 integer such that this holds which is :

$$m = \left\lceil \log_B \frac{P(R)}{T} \right\rceil \quad (1)$$

- 583 • Observe that in each pass, we read and write each page in the rela-
 584 tion exactly once. Thus, in m passes, the partitioning phase requires
 585 $2P(R)m$ IOs.
- 586 • The assumption of no skew is likely not met in practice, which is why
 587 we use statistics to estimate the skew (e.g., histograms). This is a
 588 critical point and where our simplified analysis breaks down: without

¹²If we use the same hash function, the bucket size would not reduce.

589 this hash join looks better than sort merge join in *nearly* all cases in
 590 our simple analysis. (A caveat is if we need the data in sorted order for
 591 a downstream operation, e.g., an `ORDER BY`, or if the data are already
 592 sorted, e.g., the data are stored in a clustered index.)

- 593 • We create a separate function `PartitionTimes(R, A, \bar{h}, m)` in which
 594 we replace the termination condition of the while loop of Algorithm 3
 595 with simply running m times. In this case, we produce B^m buckets.

Data: Two relations R and S both having attribute A
Result: The Join of $R \bowtie S$ on A
 $\text{Buckets_of_R} \leftarrow \text{Partition}(R, A, \bar{h}, B - 1);$
% Each bucket of R occupies B - 1 pages;
 Suppose $|\text{Buckets_of_R}| = B^m;$
 $\text{Buckets_of_S} \leftarrow \text{PartitionTimes}(S, A, \bar{h}, m);$
foreach *Bucket* B_R **in** Buckets_of_R **do**
 | Let B_S be the corresponding **joinable** bucket to B_R ;
 | $\text{BNLJ}(B_R, B_S, B)$. *% Runs in 1 pass, since B_R fits in memory!*
end

Algorithm 4: Hash Join

596 5.4.2 Join Phase

597 The hash join algorithm works as follows. Given two relations R and S ,
 598 we partition both of them until the buckets for the smaller relation fit in
 599 memory. As we will see, *only the buckets for the smaller relation need to*
 600 *fit in memory to run the join.* Then we perform (effectively) a block nested
 601 loop join in which the entire smaller relation fits in memory.

- 602 • First, partition each R and S on the join key A until the buckets
 603 produced by *either* R or S is “small”. Below, we will see that “small”
 604 means $B - 1$ given $B + 1$ buffer pages. Let m be the number of passes.
- 605 • For each of the m passes, we use the same function for R and S . That
 606 is in phase i , we use h_i for *both* relations. This ensures that if two
 607 tuples agree on the join key, then they will be in the *same bucket* after
 608 partitioning.

609 Now the join phase. We have partitioned R into buckets and S into buckets,
 610 and now we want to compute the output.

- 611 • The key property of the partitioning is that since we use the same
 612 family of hash functions on each pass, we can select pairs of buckets
 613 B_R from R and B_S from S such that for any tuples $r \in B_R$ and $s \in B_S$,
 614 we have

$$h_i(r[A]) = h_i(s[A]) \text{ for each } i = 1, \dots, m.$$

615 Notice that this holds for exactly one pair of buckets. We call such
 616 pairs of buckets **joinable**. In particular, *if r and s can join, then they*
 617 *are in joinable buckets.* Not all tuples will have the same value of A
 618 in a bucket, so not all tuples in the buckets participate in a join.

619 • For each such pair of joinable buckets, we can run BNLJ. More directly,
 620 we compute B_R joined with B_S . Note that we use our $B + 1$ buffer
 621 pages as follows:

- 622 – $B - 1$ pages for B_R
- 623 – 1 page to hold a single page B_S at a time.
- 624 – 1 page to hold the output.

625 Since we have $|B_R| \leq B - 1$ from the partitioning phase, we hold
 626 the whole block B_R in memory and scan through B_S one page at a
 627 time. We then compute whether each pair of tuples in these blocks
 628 match. In IOs, this costs the number of pages in B_R plus the number
 629 of pages in B_S . Since these are partitions and joinable buckets in a
 630 $1 - 1$ relationship, the total cost of this phase is $P(R) + P(S) + \text{OUT}$.

631 **Cost** Assume $P(R) \leq P(S)$ w.l.o.g., and that we have $B + 1$ buffer pages.
 632 The goal is to partition the data until the buckets of R (the smaller relation)
 633 are less than size $B - 1$ (i.e., take $T = B - 1$ in Equation 1). By our
 634 partitioning phase above, we require m passes where:

$$m = \lceil \log_B \left\lceil \frac{P(R)}{B - 1} \right\rceil \rceil$$

635 Recall that in each pass, we read and write each page in each relation ex-
 636 actly once. Thus, the entire partitioning phase costs $2m(P(R) + P(S))$ IOs.
 637 Finally, we need to do one final pass over the data for the join phase, which
 638 takes a single pass. Thus, our total cost in IOs is:

$$\underbrace{2 \lceil \log_B \left\lceil \frac{P(R)}{B - 1} \right\rceil \rceil (P(R) + P(S))}_{\text{Partitioning Steps}} + \underbrace{P(R) + P(S)}_{\text{Final Merge Step}} + \text{OUT}$$

639 5.5 Join Algorithm Comparison

- 640 • The memory requirement to complete in m passes is proportional to
641 the number of pages in the larger relation in sorting (since both rela-
642 tions must be sorted). In contrast, in hashing the memory requirement
643 is proportional to the size of the minimum relation.
- 644 • Skew is a major challenge for hashing. In the worst case, everything
645 hashes to a single bucket! This is the point of the histogram and
646 ANALYZE questions.

647 6 Histograms

648 We discuss equiwidth and equidepth histograms in Lecture 17, and how to
649 use these structures to make estimates of the number of tuples in queries.
650 **Make sure you understand these structures!**

- 651 • The cost formulas above rely on things like skew- we need to estimate
652 that. If we have more than one join, we need to estimate its output
653 size.
- 654 • You needed to know about ANALYZE. This is probably one of the most
655 practically useful things we teach you in this course: *if your query*
656 *seems slow, make sure it has the best available statistics.*
- 657 • Maintaining accurate statistics seems easy but is fairly tricky: we want
658 to do so in the smallest amount of space we can!
- 659 • Estimation with histograms, this is done in Lecture 17 and you should
660 know how to do it.

661 **Example 6.1 (Estimating Join Size with a Histogram)** *Consider the*
662 *join $R(A, B) \bowtie S(A, C)$. Consider a pair of buckets for the values of A in*
663 *R and the values of A in S .*

- 664 • *Suppose one bucket for $R[A]$ spans 5 values with frequency 10, we*
665 *estimate that each value occurs twice ($10/5 = 2$).*
- 666 • *Suppose one bucket for $S[A]$ spans 6 values with frequency 18, we es-*
667 *timate that each value occurs three times ($18/6 = 3$).*
- 668 • *Suppose these buckets overlap on 4 values on A , then we estimate*
669 *$(2 * 3) * 4 = 24$ values produced by this pair of buckets in the join.*

670 To obtain an estimate for the entire output, we sum the contributions of each
 671 pair of buckets to produce the estimate for the join. However, some pairs of
 672 buckets will not overlap at all, and so will contribute 0 to the output.

673 An alternate strategy, is that we could also write it like this:

$$|R(A, B) \bowtie S(A, C)| = \sum_{a \in \text{Dom}(A)} |\sigma_{A=a} R| |\sigma_{A=a} S|$$

674 here $\text{Dom}(A)$ is the domain of values for A . Using our standard histogram
 675 algorithm from class, we can estimate $|\sigma_{A=a} R|$ for any value a .

676 The first algorithm above just uses the buckets, which may be more effi-
 677 cient if the buckets are large (which is not a major concern). The reason to
 678 introduce it is to make sure you know how to operate on the buckets them-
 679 selves.