Column Databases
Exercise

• Tables:
  – Sales(Date,FKShop,FKCust,FKProd,UnitPrice,Q,TotPrice)
  – Shops(PKShop,Name,City,Region,State)
  – Customer(PKCust,Nome,FamName,City,Region,State,Income)
  – Products(PKProd,Name,SubCategory,Category,Price)
Exercise

• Sales: NRec: 100.000.000, Npag: 1.000.000; Shops: 500, 2; Customers: 100.000, 1.000; Products: 10.000, 100

SELECT Sh.Region, Month(S.Date), Sum(TotPrice)
FROM Sales S join Shops Sh on FKShops=PKShops
GROUP BY Sh.Region, Month(S.Date)

• Propose a data organization based on this query where Month(18/05/2018)=‘05/2018’ (not ‘05’)

• Consider primary organization and indexes

• Consider the possibility of denormalization and vertical partitioning
Exercise

• Compute the cost of an optimal access plan based on this organization
• Add a condition: WHERE 1/1/2017 < Date
• Assume that we want to optimize some variants as well (where the SELECT clause changes according to the GROUP BY)
  ...GROUP BY Sh.City, Year(S.Date)
  ...GROUP BY S.Date
  ...GROUP BY Sh.Region, S.FKCust
References

Column stores in 1985

- **Row store (N-ary Storage Model, NSM)**

  |
  | Page Header |
  | Id1 | John | 32 | HK245 |
  | Id3 | Mary | 33 | HK324 |
  | Id4 | John | 45 | HK245 |

- **Column store (Decomposition Storage Model, DSM)**

<table>
<thead>
<tr>
<th>Header</th>
<th>Header</th>
<th>Header</th>
<th>Header</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Id1</td>
<td>1 John</td>
<td>1 32</td>
<td>1 HK245</td>
</tr>
<tr>
<td>2 Id3</td>
<td>2 Mary</td>
<td>2 33</td>
<td>2 HK324</td>
</tr>
<tr>
<td>3 Id4</td>
<td>3 John</td>
<td>3 45</td>
<td>3 HK245</td>
</tr>
</tbody>
</table>
The reasons behind current trend

• Applicative:
  – Diffusion of analytical tasks

• Technological:
  – Widening the RAM – I/O time gap: must access disk less
  – Widening the seek-transfer gap: disk access must be sequential
  – RAM is bigger: I/O is not the only concern any more
  – Widening the cache – RAM time gap: must reduce the cache miss
  – Instruction pipelining: must reduce function call
  – SIMD instruction: make better use of SIMD parallelism
Column stores: pro and cons

• The advantages of column stores
  – Read the relevant columns only
  – Columns are easier to compress than rows: less I/O
  – Moving to tuple-at-a-time to block-at-a-time:
    • Better use of cache, pipelining, SIMD

• The problems with column stores
  – Tuple reconstruction requires a join
  – Tuple insertion requires many I/O operations
Column store performance

- From [Abadi et al. 2012], like many other pictures
How to sort a column

• Columns stored in RID order:
  – Tuple reconstruction by sort-merge-like algorithm, parallel scanning
  – No need to store the RID
• Column stored in value order:
  – Excellent compression by Run-Length-Encoding
  – Range search with optimal efficiency
  – Needs to store the RID in some way
  – Tuple reconstruction by index-nested-loop like random access
  – The column looks like an index
# Storing a column: the RID

**Sorted by RID**

<table>
<thead>
<tr>
<th>RID-value</th>
<th>One RID per page</th>
<th>Implicit RID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 20 2 50</td>
<td>1 20 50 45</td>
<td>20 50 45 20</td>
</tr>
<tr>
<td>3 45 4 20</td>
<td>20 45 20 34</td>
<td>45 20 34 50</td>
</tr>
<tr>
<td>5 45 6 20</td>
<td>8 50</td>
<td></td>
</tr>
<tr>
<td>7 34 8 50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Sorted by value**

<table>
<thead>
<tr>
<th>Value-RIDS</th>
<th>Column + Join index</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 1 4 6</td>
<td>20 *3 34 *1</td>
</tr>
<tr>
<td>34 7 45 3</td>
<td>45 *2 50 *2</td>
</tr>
<tr>
<td>5 50 2 8</td>
<td></td>
</tr>
</tbody>
</table>

1. ** RID-value**: The RIDs are stored in sorted order by RID value.
2. **One RID per page**: Each page contains one RID per column.
3. **Implicit RID**: RIDs are implicitly assigned based on the column values.
4. **Value-RIDS**: The RIDs are stored in sorted order by the corresponding value.
5. **Column + Join index**: The column values and corresponding RIDs are stored together with an index for joins.
C-Store: projections

- **A projection** is a set of columns stored together and sorted according to one of them.
- Optimal for $\sigma_{k_1 < \text{date} < k_2} \pi_{\text{saleid}, \text{date}, \text{region}}(\text{Sales})$
Projections

• Storing (saleid, date, region | date)

• Advantages of sorting date by date:
  – Better compression
  – Optimal I/O for range queries on date
  – No cost for group-by on date

• Advantages of sorting saleid, region by date:
  – No cost for join with similarly sorted columns
  – Optimal I/O for range queries on date
  – No cost for group-by on date
What is a projection

- It is not storing half-rows:
  - Every column is stored as a column
- A single sorted column is like a compressed index
- A projection (saleid, date, region | date) is like a compressed index on date with the additional attributed saleid and region
- If a column is not RID-sorted, then RIDs should be store, either in the column or in a Join Index
- Can we avoid that?
Vectorized execution

• Traditional dicotomy:
  – Tuple-at-a-type (tuple pipeline) execution: too many function calls, does not exploit array parallelism
  – Full execution of each operator: much faster on modern CPUs but intermediate results may not fit main memory

• Solution: each ‘next()’ call returns a block of values – typically, 1000 – that fits L1 cache size

• This reduces 1000 times the number of next() call

• A tight loop on a 1000 elements array can be SIMD parallelized and subject to parallel cache load
Compression

• Trading I/O – or even memory access - vs CPU
• Bit saving means more values in the registers
• Importance of fixed-width arrays
Compression algorithms

- **RLE (Run Length Encoding):**
  - a,a,a,c,c,d,c,c,c,c: (a,1,3), (c,4,2), (d,6,1), (c,7,4)

- **Bit-Vector encoding:**
  - a,b,a,a,b,c,b: a:1011000, b:0100101, c:0000010

- **Dictionary encoding:**
  - john,mary,john,john: (0:john)(1:mary) – 0,1,0,0
Compression algorithms

• Frame Of Reference
  – 1003,1017,1005: 1010 + (-7, +7, -5)

• Difference encoding
  – 1003,1017,1005: 1003, +14, -12

• Frequency partitioning:
  – Put similar values in the same page
  – Separate dictionary per page

• The patching technique: see the book if curious
Operating on compressed data

- RLE encoded data can be operated upon without decompression
- Monotone dictionary encoding allows range queries
- Bit-Vectory encoding allows bit operations on sets
Tuple reconstruction

• Early materialization:
  – ScanTable+Project(A,B,C) ->
    ScanColumn(A)+ScanColumn(B)+Join+ScanColumn(C)+Join

• Late materialization: Column Algebras

• Multi-column blocks
Storage formats

- Pure columns
- Multi-column blocks
- Column groups
Binary algebras: MIL

- XML primitives for querying a fragmented world, PA Boncz, ML Kersten, the VLDB Journal 8 (2), 101-119
- Every table is binary (or unary as a special case)
- `select(bat[H,T] AB, bool *f,...pi...) : bat[H,nil]`
  \[= \langle [a,nil] \mid [a,b] \in AB \land (*f)(b,...pi...) \rangle \]
- `join(bat[T1,T2] AB, bat[T2,T3] CD, bool *f,...pi...) : bat[T1,T3]`
  \[= \langle [a,d] \mid [a,b] \in AB \land [c,d] \in AB \land (*f)(b,c,...pi...) \rangle \]
- `reconstruct(bat[H,nil] AN, bat[H,T] BC) : bat[H,T]`
  \[= \langle [a,b] \mid [a,b] \in AB \land [a,nil] \in AN \rangle \]
Binary algebras

• *reverse*(*bat*[H,T] AB): bat[T,H]
  
  = ⟨ [b,a] | [a,b] ∈ AB ⟩

• *voidtail*(*bat*[H,T] AB): bat[H,nil] = ⟨ [a,nil] | [a,b] ∈ AB ⟩

• *group*{[a,A],[b,B],[c,A],[d,B]} = { [a,a],[b,b],[c,a],[d,b] }

• *group*(*bat*[oid,T] AB): *bat*[oid,oid]
  
  = { [a,o] | o = id_AB(b) ∧ [a,b] ∈ AB }

  where id_AB is a bijection and [id_AB(x),x] ∈ AB

• *sum*(*bat*[oid,Int] AB): *bat*[oid,Int]
  
  = [nil,sum{i | [o,i] ∈ AB }]

SELECT sum(R.a) 
FROM R, S 
WHERE R.c = S.b AND 5 < R.a < 10 
AND 30 < R.b < 40 AND 55 < S.a < 65

\( \sum(\pi_{R.a} ( \sigma_{5 < R.a < 10 \text{ and } 30 < R.b < 40} R \ join_{R.c = S.b} \ \sigma_{55 < S.a < 65} S)) \)
\[
\text{sum}(\pi_{R.a} (\sigma_{5<R.a<10 \text{ and } 30<R.b<40} R \ join_{R.c=S.b} \sigma_{55<S.a<65} S))
\]

**Initial Status**

<table>
<thead>
<tr>
<th>Relation R</th>
<th>Relation S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ra</td>
<td>Sa</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>16</td>
<td>49</td>
</tr>
<tr>
<td>56</td>
<td>58</td>
</tr>
<tr>
<td>9</td>
<td>49</td>
</tr>
<tr>
<td>11</td>
<td>49</td>
</tr>
<tr>
<td>27</td>
<td>58</td>
</tr>
<tr>
<td>8</td>
<td>97</td>
</tr>
<tr>
<td>41</td>
<td>75</td>
</tr>
<tr>
<td>19</td>
<td>42</td>
</tr>
<tr>
<td>35</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

**Query and Query Plan (MAL Algebra)**

- Inter1 = \(\sigma_{5<R.a<10} \ Ra\)
- Inter2 = \(Rb \ \text{semijoin} \ \sigma_{5<R.a<10} \ Ra\)
- Inter3 = \(\sigma_{30<R.b<40} \ Inter2\)
- join_input_R = \(Rc \ \text{semijoin} \ Inter3\)

1. Inter1 = select(Ra, 5, 20)
2. Inter2 = reconstruct(Rb, inter1)
3. Inter3 = select(inter2, 30, 40)
4. join_input_R = reconstruct(Rc, inter3)
5. Inter4 = select(Sa, 55, 65)
6. Inter5 = reconstruct(Sb, inter4)
7. join_input_S = reverse(inter5)
8. join_res_R_S = join(join_input_R, join_input_S)
9. Inter6 = voidTail(join_res_R_S)
10. Inter7 = reconstruct(Ra, inter6)
11. result = sum(inter7)
\[ \text{sum}(\pi_{R.a} (\sigma_{5<R.a<10 \text{ and } 30<R.b<40} R \ join \ R.c=S.b \ \sigma_{55<S.a<65} S)) \]

### Initial Status

<table>
<thead>
<tr>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ra</td>
<td>Sa</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>16</td>
<td>49</td>
</tr>
<tr>
<td>56</td>
<td>99</td>
</tr>
<tr>
<td>9</td>
<td>37</td>
</tr>
<tr>
<td>11</td>
<td>53</td>
</tr>
<tr>
<td>27</td>
<td>64</td>
</tr>
<tr>
<td>8</td>
<td>37</td>
</tr>
<tr>
<td>41</td>
<td>53</td>
</tr>
<tr>
<td>19</td>
<td>32</td>
</tr>
<tr>
<td>35</td>
<td>50</td>
</tr>
</tbody>
</table>

### Query and Query Plan (MAL Algebra)

1. inter1 = select(Ra, 5, 20)
2. inter2 = reconstruct(Rb, inter1)
3. inter3 = select(inter2, 30, 40)
4. join_input_R = reconstruct(Rc, inter3)
5. inter4 = select(Sa, 55, 65)
6. inter5 = reconstruct(Sb, inter4)
7. join_input_S = reverse(inter5)
8. join_res_R_S = join(join_input_R, join_input_S)
9. inter6 = voidTail(join_res_R_S)
10. inter7 = reconstruct(Ra, inter6)
11. result = sum(inter7)

### Inter4

\[ \text{Inter4} = \sigma_{55<S.a<65} S_a \]

### Inter5

\[ \text{Inter5} = \text{Rb semijoin} \ \text{Inter4} \]

### join_input_S

\[ \text{join_input_S} = \pi_{S.b} \ \sigma_{55<S.a<65} S \]
8. join_res = $(\pi_{R.c} \sigma_{5<R.a<10 \text{ and } 30<R.b<40} R \ join_{R.c=S.b} \sigma_{55<S.a<65} S))$

9. Inter6 = $\pi_{R.*} (\sigma_{5<R.a<10 \text{ and } 30<R.b<40} R \ join_{R.c=S.b} \sigma_{55<S.a<65} S)$

10. Inter7 = $\pi_{R.a} (\sigma_{5<R.a<10 \text{ and } 30<R.b<40} R \ join_{R.c=S.b} \sigma_{55<S.a<65} S)$

11. Result = $\sum(\pi_{R.a} (\sigma_{5<R.a<10 \text{ and } 30<R.b<40} R \ join_{R.c=S.b} \sigma_{55<S.a<65} S))$
Column join

• Column stores avoid indexes, hence:
  – Hash Join
  – Sort-merge
  – Main-memory join (column often fit main memory)

• Join Index: value columns -> position pairs
The Jive join

The Jive join: using the join index

<table>
<thead>
<tr>
<th>Student</th>
<th>Course</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smith(^1)</td>
<td>101</td>
</tr>
<tr>
<td>Smith(^2)</td>
<td>109</td>
</tr>
<tr>
<td>Jones</td>
<td>104</td>
</tr>
<tr>
<td>Davis(^1)</td>
<td>102</td>
</tr>
<tr>
<td>Davis(^2)</td>
<td>105</td>
</tr>
<tr>
<td>Davis(^3)</td>
<td>106</td>
</tr>
<tr>
<td>Brown</td>
<td>102</td>
</tr>
<tr>
<td>Black</td>
<td>103</td>
</tr>
<tr>
<td>Frick</td>
<td>107</td>
</tr>
</tbody>
</table>

Relation **Student**

<table>
<thead>
<tr>
<th>Course</th>
<th>Instructor</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>Green</td>
</tr>
<tr>
<td>102</td>
<td>Yellow</td>
</tr>
<tr>
<td>103</td>
<td>Green</td>
</tr>
<tr>
<td>104</td>
<td>White</td>
</tr>
<tr>
<td>105</td>
<td>Evans</td>
</tr>
<tr>
<td>106</td>
<td>Alberts</td>
</tr>
<tr>
<td>106</td>
<td>Beige</td>
</tr>
<tr>
<td>108</td>
<td>Red</td>
</tr>
<tr>
<td>109</td>
<td>Grey</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Student</th>
<th>Course</th>
<th>Instructor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smith(^1)</td>
<td>101</td>
<td>Green</td>
</tr>
<tr>
<td>Smith(^2)</td>
<td>109</td>
<td>Grey</td>
</tr>
<tr>
<td>Jones</td>
<td>104</td>
<td>White</td>
</tr>
<tr>
<td>Davis(^1)</td>
<td>102</td>
<td>Yellow</td>
</tr>
<tr>
<td>Davis(^2)</td>
<td>105</td>
<td>Evans</td>
</tr>
<tr>
<td>Davis(^3)</td>
<td>106</td>
<td>Alberts</td>
</tr>
<tr>
<td>Brown</td>
<td>106</td>
<td>Beige</td>
</tr>
<tr>
<td>Black</td>
<td>103</td>
<td>Green</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Student tuple id</th>
<th>Course tuple id</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
</tbody>
</table>

Join Result

Join Index

\(JR_1(a)\) Temp(a)

\(JR_1(b)\) Temp(b)

\(JR_1(c)\) Temp(c)

\(JR_2(a)\)

\(JR_2(b)\)

\(JR_2(c)\)
The Jive join

(a) $id < 3$
- Smith$^1$ 1
- Davis$^1$ 2
- Brown 2

$JR_1(a)$  Temp(a)

(b) $3 \leq id < 6$
- Jones 4
- Davis$^2$ 5
- Black 3

$JR_1(b)$  Temp(b)

(c) $6 \leq id$
- Smith$^2$ 9
- Davis$^3$ 6
- Davis$^3$ 7

$JR_1(c)$  Temp(c)

(a) $id < 3$
- 1
- 2

$JR_2(a)$

(b) $3 \leq id < 6$
- 4
- 5
- 3

$JR_2(b)$

(c) $6 \leq id$
- 9
- 6
- 7

$JR_2(c)$

<table>
<thead>
<tr>
<th>Student</th>
<th>101</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smith$^1$</td>
<td>102</td>
<td>Yellow</td>
</tr>
<tr>
<td>Davis$^1$</td>
<td>102</td>
<td>Yellow</td>
</tr>
</tbody>
</table>

$JR_2(a)$

<table>
<thead>
<tr>
<th>Course</th>
<th>Instructor</th>
</tr>
</thead>
<tbody>
<tr>
<td>104</td>
<td>White</td>
</tr>
<tr>
<td>105</td>
<td>Evans</td>
</tr>
<tr>
<td>103</td>
<td>Green</td>
</tr>
</tbody>
</table>

$JR_2(b)$

<table>
<thead>
<tr>
<th>Student</th>
<th>109</th>
<th>Grey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smith$^2$</td>
<td>106</td>
<td>Alberts</td>
</tr>
<tr>
<td>Davis$^3$</td>
<td>106</td>
<td>Beige</td>
</tr>
</tbody>
</table>

$JR_2(c)$
The Jive join: assumptions

- $R(R_{Id}, A, B, \ldots)$ and $S(S_{Id}, C, D, \ldots)$
- Want to compute $\pi_{ABCD} (R \bowtie_{A=C} S)$
- We have the JoinIndex $JI = \pi_{R_{Id}, S_{d}} (R \bowtie_{A=C} S)$
- We have $2*\text{NPag} < B*B$, but $\text{NPag}$ only includes pages in the projected (and semijoined) $S$ plus the pages of the Join Index:
  - $2*(\text{NPag} (\pi_{CD} (R \bowtie_{A=C} S)) + \text{Npag}(JI)) < B*B$
The Jive join

• We partition projected $\pi_{S_{ld,C,D}} S$ in K files $S_i$ ($K = B/2$)
• Partition is logical!
• We scan R and JI in parallel. We use 2*K buffers to create 2*K files that contain:
  - $JR_i = \pi_{AB} \left( R \bowtie_{A=C} S_i \right)$ (the K output files)
  - $JS_{ld,i} = \pi_{S_{ld}} \left( R \bowtie_{A=C} S_i \right)$ (the K temporary files)
• We are almost done: we must just create, for each $JR_i$, a corresponding $JS_i$ that contains $\pi_{CD} \left( R \bowtie_{A=C} S_i \right)$ in the same order
• Cost of the phase?
The Jive join

- For each i, we load the whole JSId_i in memory, and sort a copy in SId order
- We read the corresponding i-partition of $\pi_{SId,C,D} S$, skipping the useless blocks
- We scan the unsorted copy of JSId. For each Sid in JSId, we put the corresponding record of $S_i$ (which is in memory) in JS_i
- Now we have:
  - $JR_i = \pi_{AB} (R \bowtie_{A=C} S_i)$
  - $JS_i = \pi_{CD} (R \bowtie_{A=C} S_i)$
Memory needs

• Phase 1: $2*K$ buffers, hence $K = B/2$ ($y = m/2$)
• Phase 2:
  – Size of JSIdi: $|JI|/K$
  – Plus size of $\pi_{Sld,C,D} S$ – $Sld$ may be implicit: $|\pi_S|/K$
• Hence
  – $(|JI| + |\pi_S|) / K < B$
  – $2*|JSIdi| + 2*|\pi_S| < B*B$
Group by and aggregation

- Group by is typically hash-based
- Aggregation (sum, count...) takes advantage of columnar layout in order to perform a tight loop over a cache-sized array
Insert / update / delete

• Insertion is expensive:
  – One insert for each column
  – Redundant representation implies redundant insertions
• If a column is ordered, then insertion must respect the order
• Decompression / recompression is often needed
• Updates and deletes have the same problems
• Solution: differential files
Read Store and Write Store

• Mature data live in the big Read Store and fresh Updates (insertions/deletions/updates) are in the small main-memory Write Store
• Read Store is read-optimized while Write Store is compact
• Every query queries both the RS and the WS and merges the two results – the WS results contains insertions to be added and deletions to be removed
• Problem: this approach requires an explicit RID
RS / WS and transactions

• Implementing Snapshot Isolation:
  – The RS is the Snapshot
  – Each transaction has its own WS

• Implementing no-undo / no-log concurrency control:
  – The WS is merged only after commit
  – The WS is the redo log
Indexing in column stores

• A sorted column A with a join index corresponds to a traditional index A

• A projection \{ A, B, C \mid C \} may be considered as an index on C that allows one to rapidly access A and B, with an IndexOnly plan
Simulating columns with indexes

• We create an index on every column of R(A, B, C, D, ...)
• Would typically not work:
  – Tuple reconstruction requires a Join of the index with the original table
  – Insert/delete overhead
Simulating columns: vertical partitioning

• We may rewrite the table $R(A,B,C,D,...)$ as a set of tables $R(IdA,A)$, $R(IdA,B)$, $R(IdA,C)$, where $IdA$ is a small integer, and the tables are $IdA$-sorted for a fast join.

• However:
  – No compression
  – $IdA$ overhead
  – Pipelined execution
  – Lots of join may confuse the optimizer
  – Insert/delete overhead
Conclusions: a new landscape

• Applications: OLAP vs OLTP applications, ingesting whole tables rather than looking for a tuple
• Main memory is getting bigger and bigger
• Disk seek time is getting relatively slower
• Memory access is getting slower: needs to exploit data and code cache
• On-chip parallelism must be exploited: SIMD operations and pipelined executions: more arrays, less function calls
Conclusions: is not just vertical partitioning

- Only read the columns that you need
- Keep the columns compressed
- Store redundant projections in many orders
- Never rebuild the tuple until the end
- Use fixed-size columns and work on a cache-sized array at a time
Exercise: the cost of the following plan
SELECT sum(R.a)
FROM R, S
WHERE R.c=S.b  and  5<R.a<10
    and 30<R.b<40 and  55<S.a<65

\[ \text{sum}(\pi_{R.a} ( \sigma_{5<R.a<10 \text{ and } 30<R.b<40} R \text{ join}_{R.c=S.b} \sigma_{55<S.a<65} S)) \]