Insight2014

The Conference for Big Data and Analytics

IBM DB2 11 for z/OS Optimizer Update
Session: IDZ-5363

Michelle Guo

Terry Purcell

IBM DB2 Development

Delivered to NEDB2UG by Mark Rader

March 26, 2015

#ibminsight

Agenda

- **Plan Management**
- **Predicate Indexability**
- **In-Memory Data Cache (sparse index)**
- **Duplicate Removal**
- **DPSIs and page range**
- **Misc Performance enhancements**
- **Optimizer externalization and input enhancements**

Plan Management Enhancements

Static Plan Management – APREUSE(WARN)

- DB2 10 delivered APREUSE(ERROR)
  - Allowed potential for reuse of prior plan to generate new runtime structure
  - Failure of reuse failed the entire package
  - APREUSE(ERROR) EXPLAIN(ONLY) failure may not represent a valid plan in DB2 10
    - Failed access path is written to PLAN_TABLE
- DB2 11 delivers APREUSE(WARN)
  - Upon failure of reuse, Optimizer will generate a new access path choice
    - Thus failure of 1 SQL will not fail the entire package
  - PLAN_TABLE output will represent a valid plan
    - For both ERROR or WARN

APREUSE Comparison of ERROR and WARN

- APREUSE(ERROR)
  - Effectively operates at the package level
- APREUSE(WARN)
  - Effectively operates at the statement level
- Example
  - REBIND on a workload in which 3% of the queries fail their respective hints
    - With APREUSE(WARN)
      - Access paths kept on all statements that took the hint
      - Fresh access paths for statements on which the hint failed
      - All packages rebound successfully and 97% of SQLs succeed
    - With APREUSE(ERROR)
      - Access paths kept for all packages that took all hints
      - Package REBIND failure where a hint failed
      - Worst case 100% of packages fail (if each has ~30 SQL and 1 fails)

Predicate Indexability Improvements

Rewrite Common Stage 2 predicates to indexable

- Targeted Stage 2 predicates
  - YEAR(DATE_COL)
  - DATE(TIMESTAMP_COL)
  - value BETWEEN C1 AND C2
  - SUBSTR(C1,1,10) ← SUBSTR from position 1 only
- Stage 2 predicates ONLY rewritten if no candidate Index On Expression to support the predicate
  - Regardless of whether the optimizer chooses that IOE
- Applies to literals or host variables/parameter markers
  - Requires REBIND for static
- NOTE: Increase in matchcols will cause APREUSE(ERROR) to fail
  - APREUSE(WARN) will succeed if only change is matchcol increase

Stage 2 predicate rewrite examples

- EQUAL Example:
  - WHERE YEAR(DATE_COL) = 2012 ← stage 2

  Becomes
  - WHERE DATE_COL BETWEEN ‘2012-01-01’ AND ‘2012-12-31’ ← indexable

- Also applies to IN, BETWEEN, range predicates etc

- Range Example:
  - WHERE SUBSTR(CITY,1,3) <= :hv ← stage 2

  Becomes
  - WHERE CITY <= (exp) ← Indexable
    (exp is a DB2 computed value for boundaries of column)
    - For example: SUBSTR(CITY,1,3)<=‘ABC’
      - Becomes CITY<=x’C1C2C3FFFFFFFFFFFFFF’

Value BETWEEN two columns

- Example:
  - SELECT *
    FROM TABLE
    WHERE :hv BETWEEN START_DATE AND END_DATE ←Stage 2

    Becomes

  - SELECT *
    FROM TABLE
    WHERE START_DATE <= :hv ← indexable
    AND END_DATE >= :hv ← indexable

    NOTE: COL BETWEEN :hv1 AND :hv2 is already indexable

Indexability for OR/IN and OR COL IS NULL

- Improved single matching index access for OR C1 IS NULL
  - Examples

    WHERE C1 = ? OR C1 IS NULL

    WHERE C1 IN (1, 2) OR C1 IS NULL

    WHERE C1 > ? OR C1 IS NULL

- IN/OR combination to allow multi-index access……

  WHERE C1 = ? OR C2 IN (1,2)

  Becomes

  WHERE C1 = ? OR C2 = 1 OR C2 = 2

Prune always true predicates

- Example WHERE 1=1
  - So what's the problem with this harmless predicate?
    - DB2 will execute the WHERE 1=1 predicate for every qualified row
    - SELECT *
      FROM TABLE
      WHERE 1=1
      AND CUSTNO = ?
  - Prune always true predicate to become
    - SELECT *
      FROM TABLE
      WHERE CUSTNO = ?

Prune always false predicates

- DB2 10 already prunes “always false” equal/IN under OR
  - WHERE C1 = ? OR ‘A’ = ‘B’
- DB2 11 extends to “always false” underneath parent “AND”
  - SELECT *
    FROM TABLE1 T1, TABLE2 T2
    WHERE (1=1 AND T1.C1 = T2.C1)
    OR (1=2 AND T1.C2 = T2.C2)
  - Prune always true/false predicates to become
    - SELECT *
      FROM TABLE1 T1, TABLE2 T2
      WHERE T1.C1 = T2.C1
- NOTE: “OR 0=1” is NOT pruned
- NOTE2: Literals only. No host vars/markers. No reopt.

Indexability for CASE predicates

- Case can now be indexable (formerly stage 2)
  - For local predicate

```
– SELECT *  FROM T1
  WHERE COL =      CASE (CAST(? AS INT))
                   WHEN 1 THEN 'CA'
                   WHEN 2 THEN 'NY'
                   ELSE 'AL' END;
```

  - For JOIN predicate
    - CASE expression must be evaluated before the join.
    - In example below, join predicate is indexable if T1 accessed before T2.

```
– SELECT * FROM T1, T2
  WHERE T2.COL =   CASE WHEN T1.COL = 'Y'
                   THEN T1.COL2
                   ELSE T1.COL3
                   END;
```

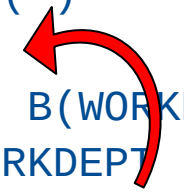

Predicate Pushdown

- DB2 11 pushdown into materialized views/Table Expressions of
  - Non-boolean term (OR) predicate

```
SELECT EMPNO, SALARY, DEPTCOUNT
FROM   EMP A ,
(SELECT WORKDEPT, COUNT(*)
 FROM   EMP
 GROUP  BY WORKDEPT) AS B(WORKDEPT, DEPTCOUNT)
WHERE A.WORKDEPT = B.WORKDEPT
AND  (B.WORKDEPT LIKE 'C%' OR B.WORKDEPT LIKE 'A%‘);
```

- Stage 2 predicates (expressions)

```
SELECT EMPNO, SALARY, DEPTCOUNT
FROM   EMP A ,
(SELECT WORKDEPT, COUNT(*)
 FROM   EMP
 GROUP  BY WORKDEPT) AS B(WORKDEPT, DEPTCOUNT)
WHERE A.WORKDEPT = B.WORKDEPT
AND UPPER( B.WORKDEPT) = 'C01'
```



Predicate Pushdowns (cont.)

- Predicate in the ON clause of an outer join

```
SELECT EMPNO, SALARY, DEPTCOUNT
FROM   EMP A
LEFT OUTER JOIN
(SELECT WORKDEPT, COUNT(*)
 FROM   EMP
 GROUP  BY WORKDEPT) AS B(WORKDEPT, DEPTCOUNT)
ON A.WORKDEPT = B.WORKDEPT
AND B.WORKDEPT = 'C01';
```

- Also
  - when the view/Table Expression contains a scalar function in the SELECT list
- Some restrictions still remain, such as:
  - If all 3 examples had predicates against table A – predicate not pushed in
  - Expression pushdown may not qualify for index on expression

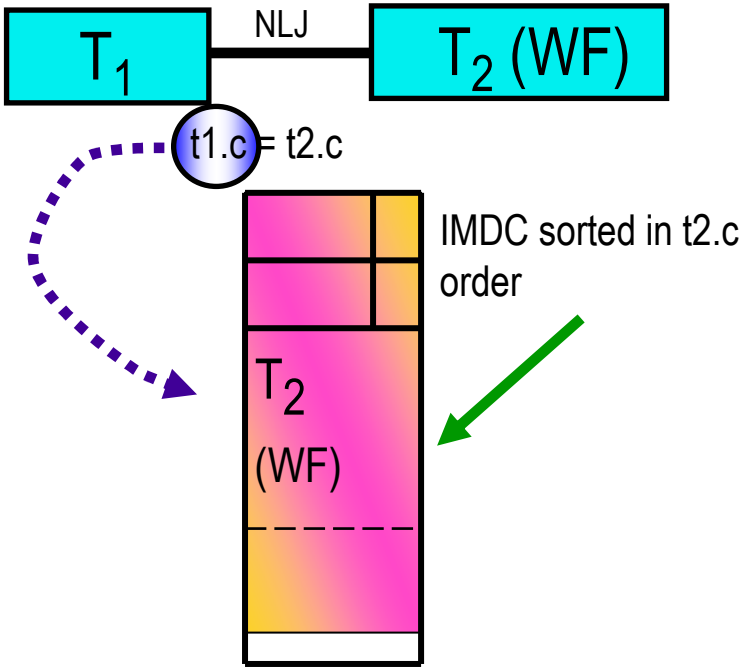
In-memory data cache / sparse index

In Memory Data Cache & Sparse Index

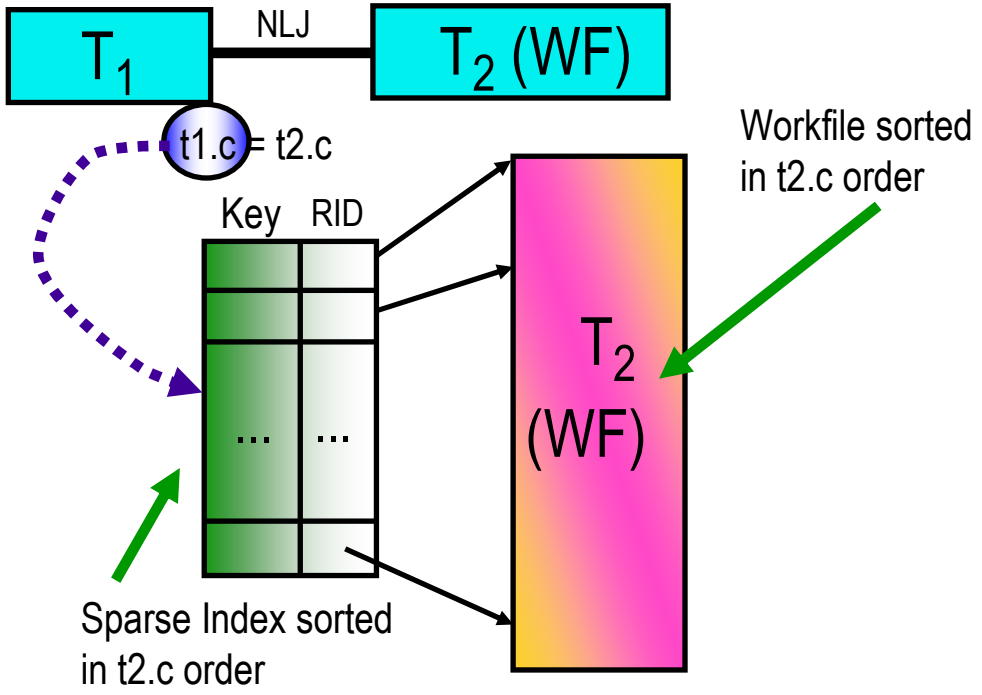
- History
  - V4: Sparse Index for non-correlated subquery workfiles
  - V7: Sparse Index for materialized workfiles for star join
  - V8: IMDC enhancement for star join
    - Fallback to sparse index when insufficient memory
  - V9: IMDC / SI extended to non-star join when table lack index on join columns (Generalized Sparse Index)
    - Also supports multicolumn sparse index
    - MXDTCACH ZParm
      - Maximum memory for data caching per thread
      - 0-512MB (default 20MB)
      - 0 = Only Sparse Index (key+rid) applied
  - V10: Add hash support for sparse index (as alternative to binary search)
    - When result would be contained in MXDTCACH

In Memory Data Cache vs. Sparse Index

- IMDC (hash or binary search)
- Sparse Index
  - When insufficient memory for IMDC



Binary Search of WF to look up exact location of qualified key (Hash used if sufficient memory)



Binary Search of sparse index to look up "approximate " location of qualified key

IMDC/Sparse Index DB2 11 Enhancements

- Improved memory management by optimizer and runtime
  - Controlled by zparm MXDTCACH (default 20MB)
  - Each sparse index/IMDC is given a % of MXDTCACH
    - From optimizer cost perspective
    - At runtime (based upon cost estimation)
  - Runtime will choose appropriate implementation based upon available storage
    - Hash, binary search, or spill over to workfile
- Improved optimizer cost model
  - Allowing this to be opened up in more join scenarios
- Improvements to IFCID 27 for detail, 2 & 3 for summary

IMDC/Sparse index – Tuning basics

- DB2 11 provides simple accounting/statistics data for sparse index
  - Sparse IX disabled
    - indicates main memory was insufficient for the MXDTCACH memory request
    - Suggest reducing MXDTCACH or allocating more memory to the system
  - Sparse IX built WF
    - MXDTCACH was insufficient to contain sparse index
      - Increase MXDTCACH
    - Look at sort BP sync I/O
      - If high, also reduce VPSEQT in sort BP (do not use VPSEQT=100)

```
MISCELLANEOUS              AVERAGE      TOTAL
-------------------------  --------  --------
SPARSE IX DISABLED            0.00         0
SPARSE IX BUILT WF            0.36         8
```


Duplicate Removal

Index skipping and Early-out – DB2 11 Enhancements

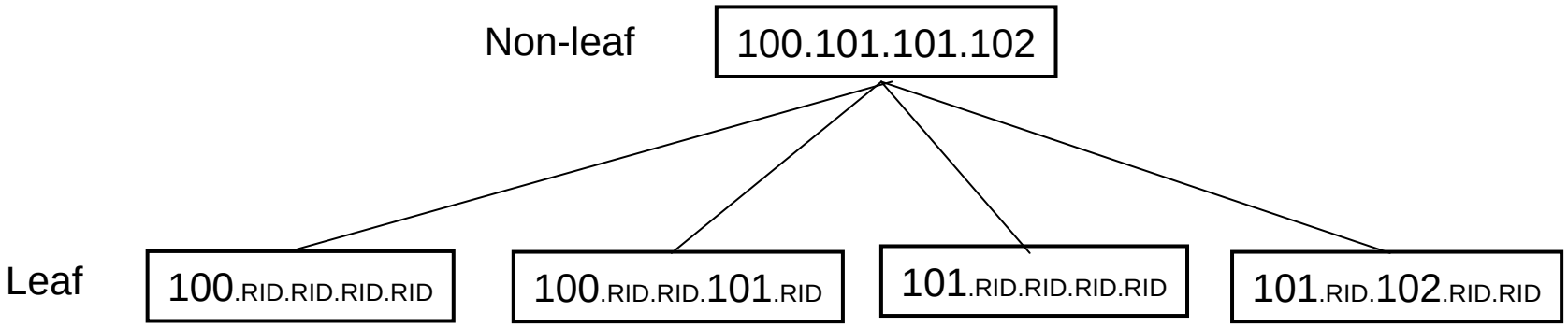
- Improvements to queries involving GROUP BY, DISTINCT or non-correlated subq
  - Where an index can be used for sort avoidance
    - By skipping over duplicates (see next few slides)

- Improvement to join queries using GROUP BY, DISTINCT
  - By NOT accessing duplicates from inner table of a join if DISTINCT/GROUP BY will remove those duplicates

- Improvement to correlated subqueries
  - Early-out for ordered access to MAX/MIN correlated subqueries
    - When I1-fetch is not available
  - Optimize usage of the “result cache” for access to subquery with duplicate keys from the outer query
    - 100 element result cache dates back to DB2 V2 as a runtime optimization
    - DB2 11 adds optimizer recognition of benefit

Pre-DB2 11 Duplicate Removal using an index (no sort)

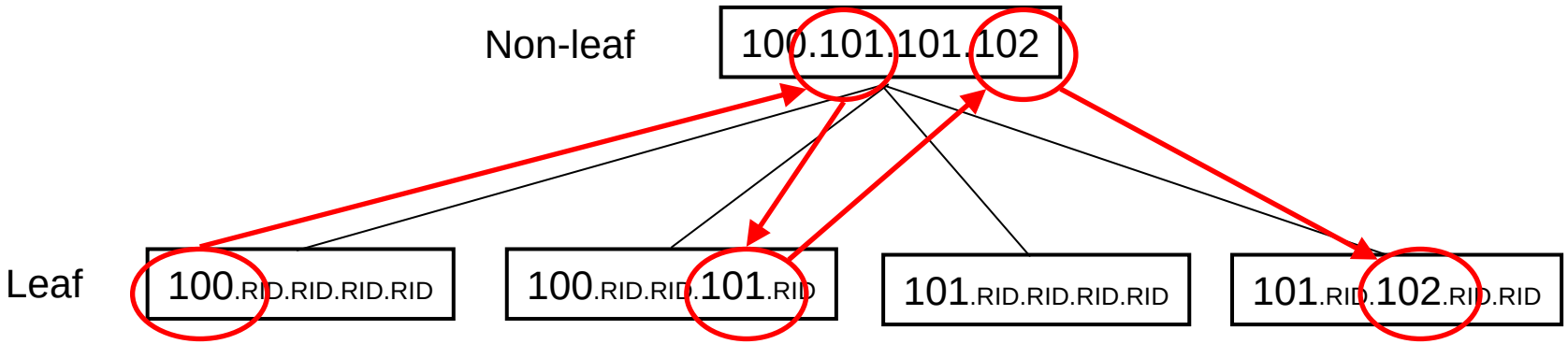
```
SELECT C1
FROM T
GROUP BY C1
```



Scan qualified leaf pages (and all rids) with runtime discarding duplicates

DB2 11 - Duplicate Removal with Index Skipping

```
SELECT C1
FROM T
GROUP BY C1
```



Index Skipping (over-simplified)

Use index lookaside (current leaf high key and non-leaf) to get the next key greater than current key

Early-out join

- DB2 11 supports early-out for joins if duplicates not necessary
  - Previously only avail for correlated EXISTS subquery transformed to join.
  - For below example: Duplicates from T2 are removed by DISTINCT
    - In DB2 11, each inner table probe will stop after 1st match is found
      - NOTE: For LEFT OUTER JOIN V10 will prune T2

```
SELECT DISTINCT T1.*
FROM T1, T2
WHERE T1.C1 = T2.C1
```

    - Also apply to Non-Boolean Term join conditions with “early-out” table

```
SELECT DISTINCT T1.*
FROM T1, T2
WHERE T1.C1 = 1
   OR T1.C1 = T2.C1
```


Optimize usage of subquery result cache

- DB2 V2 introduced a result cache for saving the 100 most recent correlated subquery execution results
  - Each subquery execution would 1st scan the cache to find the result
    - If found, cache value is used
    - If not found, subquery is executed, and result saved in cache
- DB2 11 adds optimizer recognition of the cache
  - Ordered access will reduce the cache size from 100
    - Example below, accessing the outer in CUSTNO order (via CUSTNO index or tablespace scan if CUSTNO clustering) would result in cache hits for repeat CUSTNO values

```
SELECT *
FROM POLICY P1
WHERE P1.POLICY_DATE =
(SELECT MAX(P2.POLICY_DATE)
 FROM POLICY P2
 WHERE P2.CUSTNO = P1.CUSTNO)
```


DPSI and Page Range

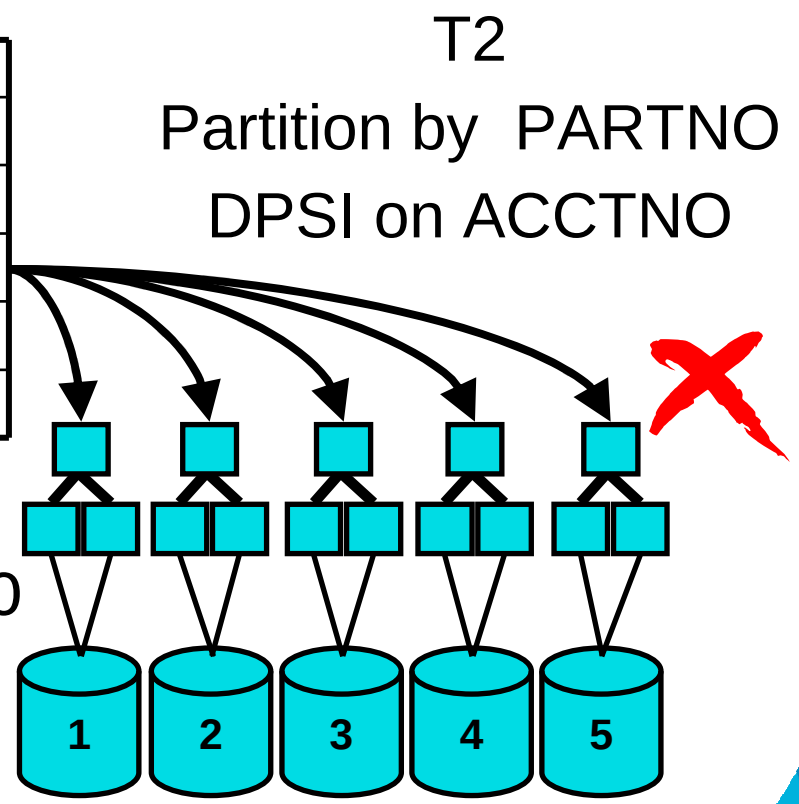
Page Range Screening – DB2 11 Enhancements

- Page range performance Improvements
  - Page Range Screening on Join Predicates
    - Access only qualified partitions
  - Pre-DB2 11, page range screening only applied to local predicates
    - With literals, host variables or parameter markers
  - Applies to index access or tablespace scan
    - Benefits NPIs by reducing data access only to qualified parts
    - Biggest benefit to DPSIs by reducing access only to qualified DPSI parts
- Only for equal predicates, same datatype/length only

Pre-V11 Page Range Join Probing (Join on partitioning Col)

- Current challenge
  - Composite row probes all parts

YEAR	PARTNO
2009	1
2010	2
2011	3
2012	4
2013	5

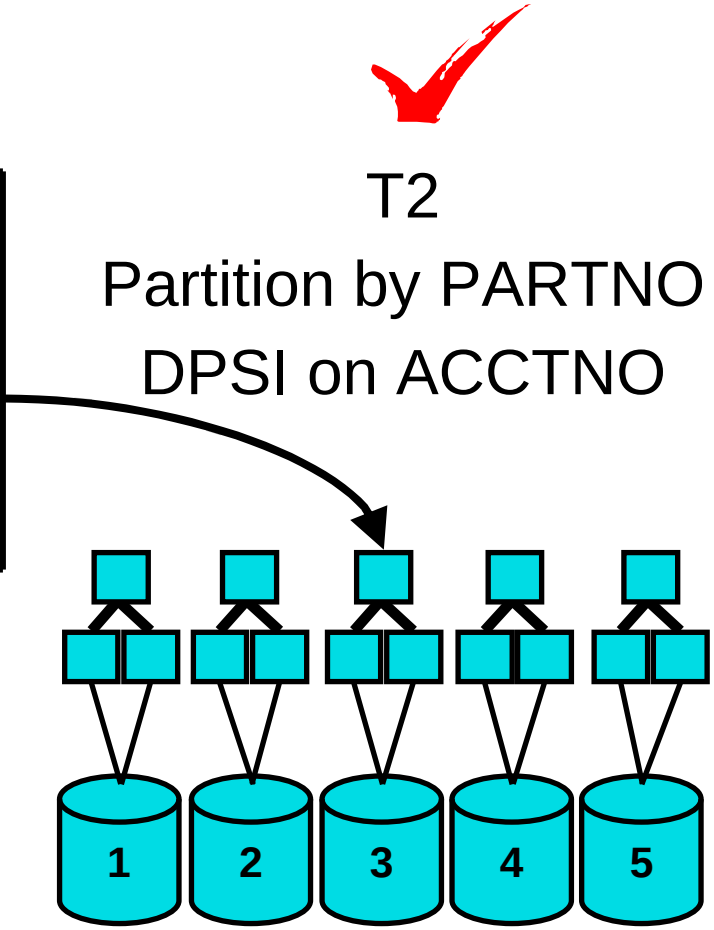



```
SELECT *
FROM T1, T2
WHERE T1.PARTNO = T2.PARTNO
AND T1.YEAR = 2011
AND T2.ACCTNO = 12345
```

V11 Page Range Join Probing (Join on Partitioning Col)

- Join recognizes page range screening
  - Only 1 partition needs probing.

YEAR	PARTNO
2009	1
2010	2
2011	3
2012	4
2013	5




```
SELECT *
FROM T1, T2
WHERE T1.PARTNO = T2.PARTNO
AND T1.YEAR = 2011
AND T2.ACCTNO = 12345
```

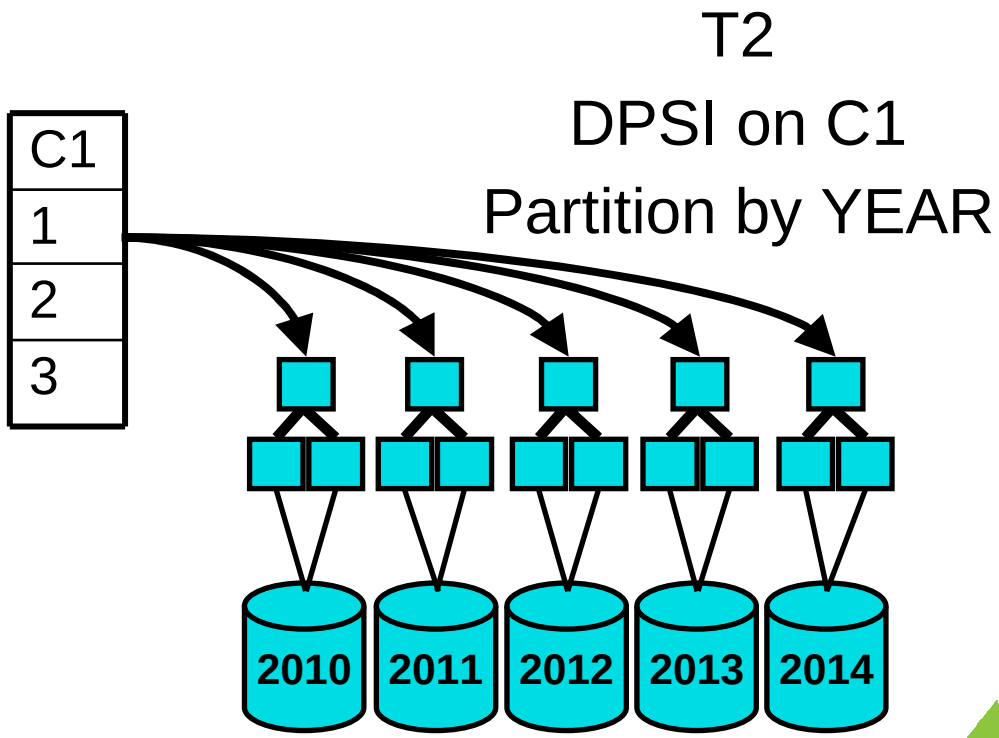
DPSI – DB2 11 Enhancements

- DPSI can benefit from page range screening from join
  - Assuming you partition by columns used in joins (see previous slides)
- For DPSIs on join columns and partition by other columns
  - DB2 11 Improves DPSI Join Performance (using parallelism)
    - Controlled by ZPARM PARAMDEG_DPSI
- Sort avoidance for DPSIs (also known as DPSI merge)
  - Use of Index On Expression (IOE)
    - Ability to avoid sorting with DPSI IOE (already available for DPSI non-IOE)
  - Index lookaside when DPSI used for sort avoidance
- Straw-model parallelism support for DPSI
  - Straw-model (delivered in V10) implies that DB2 creates more work elements than there are degrees on parallelism.

Pre-V11 DPSI Probing Challenge for Joins

- NOTE: No page range join predicate
- Current challenge for join to a DPSI
  - 1st composite row probes all parts
  - 2nd composite row probes all parts
  - Etc

```
SELECT *
FROM T1, T2
WHERE T1.C1 = T2.C1
```

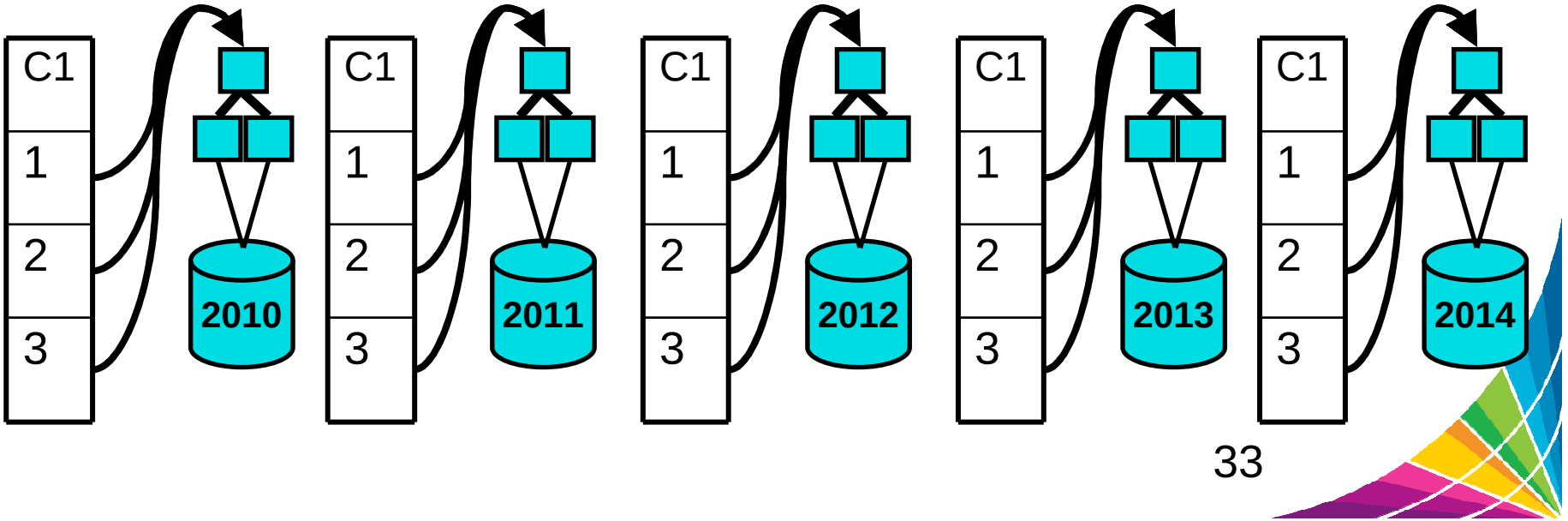


DPSI Probing – DB2 11 Join Solution

- DPSI part-level Nested Loop Join
  - Share composite table for each child task (diagram shows a copy)
    - Each child task is a 2 table join
    - Allows each join to T2 to access index sequentially (and data if high CR)

SELECT *
FROM T1, T2
WHERE T1.C1 = T2.C1

T2
DPSI on C1

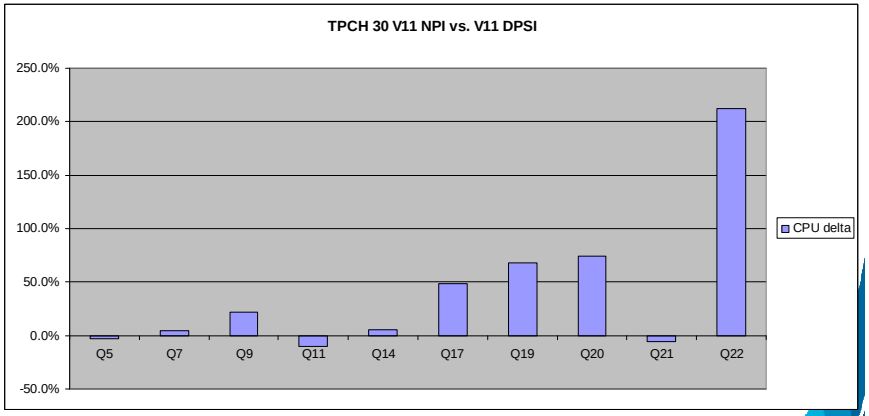


DPSI – what is the message?

- A “partitioned” index means excellent utility performance
  - But historically there was one sweet spot ONLY for DPSIs
    - When local predicates in the query could limit partitions to be accessed

- Does DB2 11 allow me to switch all NPIs to DPSIs?
  - NO, but the sweet spot just got a little bigger
    - NPIs still are necessary in many workloads

- How do NPIs & DPSIs now compare?
  - Internal TPCH measurement
    - DPSIs increased CPU on avg by 8%
      - But 1 query was 200% !!!!
  - DB2 11 ESP customer feedback
    - 2 customers reported > 75% CPU improvement for DPSIs (no other details provided)



Misc Performance Items

CPU speed impact on access paths

- DB2 11 can reduce access path changes based upon different CPUs
  - Across data sharing members
  - After CPU upgrade
  - Development vs production with different CPU speeds
  - Applies to
    - z10 to z196 or zEC12, or z196 to zEC12
      - And later CPUs

Sort / Workfile Performance

- In memory workfile support in DB2 9 and 10
  - Final sort in DB2 9 (up to 32K) and 10 (up to 1MB)
  - DB2 10 intermediate workfile usage up to 32K for selective path
- More in memory operation in DB2 11
  - Final sort up to 128MB by zparm control MAXSORT_IN_MEMORY (default 1MB)
  - Wider range of usage for in memory
    - Materialized view, table expression, outer Join, EXISTS, etc.
  - Avoid workfile usages for final merge on top level sort
  - Reduces physical workfile usage for large top level sort

- NOTE: In-memory avoided if CURSOR WITH HOLD
  - Which is the default for ODBC & JDBC

RID processing enhancements

- Pre-DB2 11
  - DB2 10 added RID failover to WF
    - Did not apply to queries involving column function
  - A single Hybrid Join query could consume 100% of the RID pool
    - Causing other concurrent queries to hit RID limit if > 1 RID block needed
- DB2 11
  - RID failover to WF extended to all scenarios when RID limit is hit
  - Hybrid join limited to 80% of the RID pool

Other interesting performance items

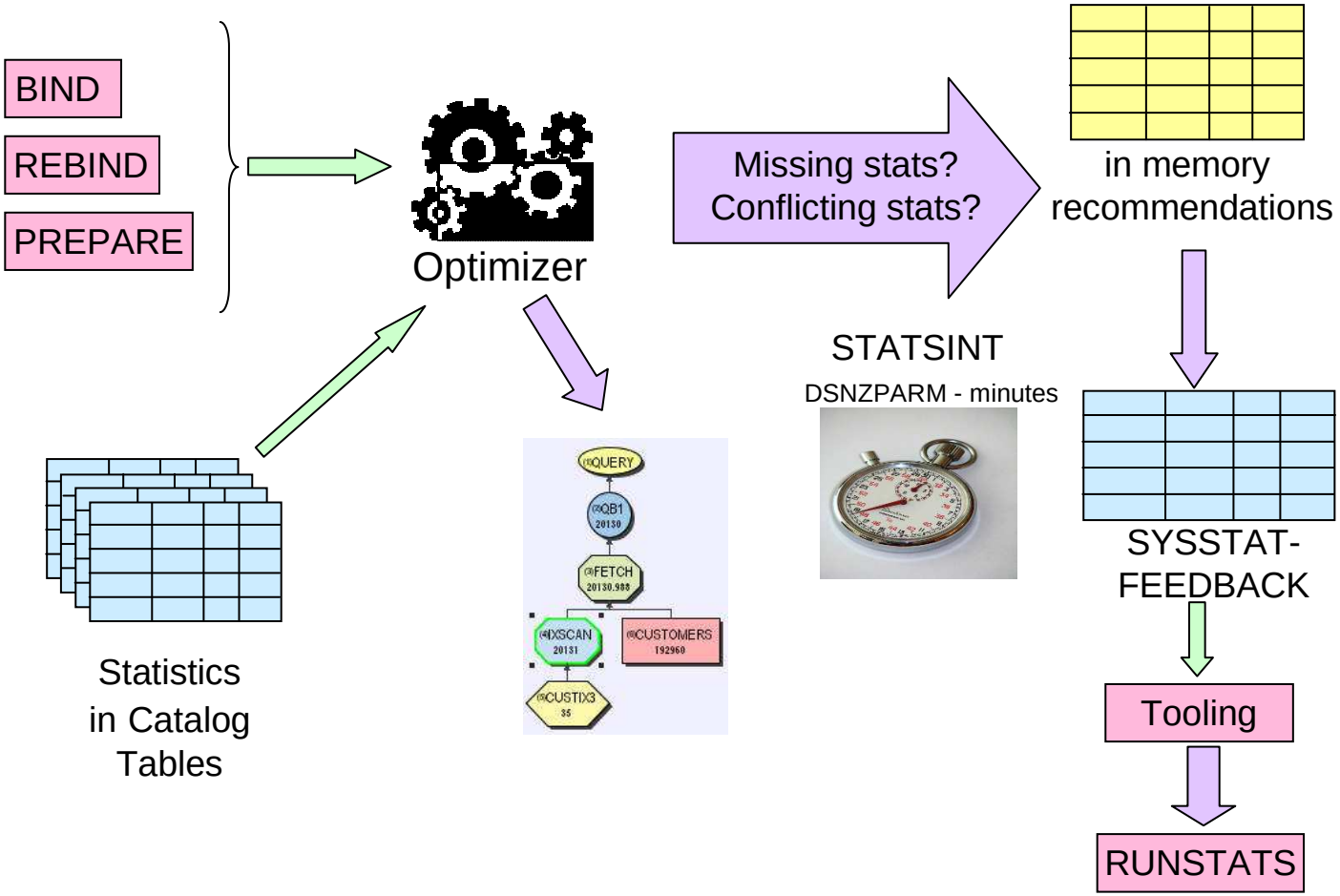
- DGTT NOT LOGGED support
- EXCLUDE NULL indexes
- Pseudo-deleted index entry cleanup
- Reduction of indirect references
- Decompression performance improvements
- DECFLOAT performance improvements (used extensively in XML)
- Tablespace scan performance improvements

Optimizer externalization of missing stats and Overriding FF estimates

DB2 Optimizer and Statistics - Challenge

- DB2 cost-based optimizer relies on statistics about tables & indexes
- Customers often gather only standard or default statistics
  - E.g. RUNSTATS TABLE(ALL) INDEX(ALL) KEYCARD
- Queries would often perform better if DB2 optimizer could exploit more complete statistics
- Customers have difficulty knowing which statistics are needed

DB2 11 – Optimizer externalization of missing statistics



DB2 11 Solution: Optimizer Externalization

- During access path calculation, optimizer will identify missing or conflicting statistics
  - On every BIND, REBIND or PREPARE
    - Asynchronously writes recommendations to SYSIBM.SYSSTATFEEDBACK
  - DB2 also provides statistics recommendations on EXPLAIN
    - Populates DSN_STAT_FEEDBACK synchronously
- Contents of SYSSTATFEEDBACK or DSN_STAT_FEEDBACK can be used to generate input to RUNSTATS
  - Contents not directly consumable by RUNSTATS
  - Requires DBA or tooling to convert to RUNSTATS input

Optimizer selectivity - The Filter Factor Problem

- Query optimization challenges
  - Cost based query optimization
    - Estimate cost of available choices to identify choice with cheapest cost
  - The optimizer needs to know how many rows are filtered at every step
    - How much does it cost to scan index ? Matching, screen filtering
    - Which table should be outer?

- Sometimes, the optimizer is unable to accurately estimate selectivity
  - Lack of statistics
  - Join skew, join correlation
  - Complex predicates
  - Predicate combinations
  - Unknowns (host variables, parameter markers, special registers)

DB2 11 Selectivity Overrides (FF hints)

- Process of supplying more robust selectivity (Filter Factor) input
  - Rather than a whole OPTHINT – just FF hints
- Selectivity profile allows User to
  - Provide optimizer with a more accurate view of selectivities used in query execution
    - For one, some or all predicates in a query
    - For one or more representative “executions” of a query
      - Weighted by frequency of occurrence
- Similar to the SELECTIVITY clause on SQL statement, but ...
    - Doesn’t require changing applications
    - Handle variations in execution
- Also has OQWT tooling support

Virtual Index Improvements

Virtual Index Enhancements – Table Changes

- DSN_VIRTUAL_INDEXES enhanced
  - Columns added to complete index modelling capabilities
    - UNIQUE_COUNT

      To support INCLUDE index columns
    - SPARSE

      To support NULL Suppressed indexes
    - DATAREPEATFACTORF

      To support enhanced statistics gathering
    - KEY_TARGET_COUNT & IX_EXTENSION

      To support Index on Expression and XML Index

- DSN_VIRTUAL_KEYTARGETS
  - New EXPLAIN table used for Index Advisor support for IOE and XML indexes

Acknowledgements and Disclaimers

Availability. References in this presentation to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates.

The workshops, sessions and materials have been prepared by IBM or the session speakers and reflect their own views. They are provided for informational purposes only, and are neither intended to, nor shall have the effect of being, legal or other guidance or advice to any participant. While efforts were made to verify the completeness and accuracy of the information contained in this presentation, it is provided AS-IS without warranty of any kind, express or implied. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this presentation or any other materials. Nothing contained in this presentation is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software.

All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer. Nothing contained in these materials is intended to, nor shall have the effect of, stating or implying that any activities undertaken by you will result in any specific sales, revenue growth or other results.

Thank You

IBM

Insight2014

The Conference for Big Data and Analytics

SEIZE THIS MOMENT

#ibminsight