PostgreSQL Bug Report

## Overview

When trying to run a complex join of eight tables – including 6 foreign tables – the query runs for a few seconds and then terminates with an error message

This occurs ONLY in the Dev environment; the query executes successfully in our Production environment, although it returns zero rows.

## The query

Here is the query that is causing the problem:

SELECT

dk.deployment\_id,

dk.deployment\_key\_txt,

dp.name as deployment\_nm,

dp.tenant as deployment\_tenant\_nm,

infc.name as deployment\_cluster\_nm,

icv.description as cluster\_kubernetes\_version\_id,

infc.nodes as cluster\_node\_no,

infc.ram as cluster\_ram\_no,

infc.cpu as cluster\_cpu\_no,

infc.disk as cluster\_disk\_no,

dp.deploy\_description as infrastructure\_desc,

CASE

 WHEN infrq.multi\_master = 1 THEN 'Y'

 else 'N'

END

AS HA\_multi\_master\_flg,

infrq.cluster\_label as cluster\_environment\_desc,

dl.deploy\_time as deployment\_deploy\_tm,

dl.order\_number as order\_no\_id,

ob.order\_id as order\_id,

dp.server as deployment\_host\_nm,

to\_timestamp(CAST(dl."timestamp" AS double precision)) as deployment\_created\_dttm,

dl.promotion\_stage as promotion\_stage\_cd,

se.ship\_event\_nm as ship\_event\_nm,

'UDANEXT' as source\_system\_cd

FROM foreign\_udanext.deployments dp

LEFT OUTER JOIN foreign\_udanext.deployment\_log dl ON (dl.deploy\_id = dp.id)

LEFT OUTER JOIN foreign\_udanext.infrastructure\_tenant inft ON (inft.name=dp.tenant)

JOIN cqm\_meta.deployment\_key dk ON (concat\_ws('|',dl.deploy\_id, dl.task\_id,'UDANEXT') = dk.deployment\_key\_txt)

-- JOIN sel\_meta\_key dk ON concat\_ws('|',dl.deploy\_id, dl.task\_id,'UDANEXT') = dk.deployment\_key\_txt

LEFT OUTER JOIN foreign\_udanext.infrastructure\_cluster infc ON (infc.tenant\_id=inft.id)

LEFT OUTER JOIN foreign\_udanext.infrastructure\_cluster\_request infrq ON (infrq.id=infc.request\_id)

LEFT OUTER JOIN foreign\_udanext.infrastructure\_cluster\_version icv ON (infc.k8s\_version\_id=icv.id)

LEFT OUTER JOIN main.ship\_event se ON (se.ship\_event\_nm = substring(dp.shipevent from 6 for 5))

inner join main.order\_base ob on ob.order\_no = dl.order\_number

WHERE dk.deployment\_key\_txt NOT IN

 (SELECT source\_system\_deployment\_id

 FROM main.deployment\_base db)

The intent is to use this SELECT within an INSERT statement that will add the result data into a table, but first we have to have the SELECT work appropriately.

## The error

ERROR: mergejoin input data is out of order

SQL state: XX000

This error message is not documented anywhere that I can find, and it only occurs in Dev, not Prod.

## Suspected issue

We believe that this error is caused by one of two conditions:

1. Postgres is assuming an incorrect sort order on data returned from a substring function; and/or
2. The number of joins is causing intermediate data to be cached, and available cache memory is being exhausted.

## Reasons for error suspicions

For (1), the error is suspected because if the line containing the substring is commented out, along with it’s associated join, the query runs successfully.

Substring:

LEFT OUTER JOIN main.ship\_event se ON (se.ship\_event\_nm = substring(dp.shipevent from 6 for 5))

Query parameter (part of select):

se.ship\_event\_nm as ship\_event\_nm,

The reason we suspect this is first, because the query runs if this join and column selected is commented out of the query. Second, the error message, “mergejoin input data is out of order”, leads us to suspect that it is expecting the shipevent data to be in a particular order, and it is not because it is a substring that begins with the 6th character of the string in the column.

What argues against this being the issue is that fact that the query executes properly in Prod, which is running on the same version of Postgres as Dev.

For error (2), the suspicion is based on the log output. The full log will be attached to this bug report, but here’s the relevant subset of the log:

 SELECT

 dk.deployment\_id,

 dk.deployment\_key\_txt,

 dp.name as deployment\_nm,

 dp.tenant as deployment\_tenant\_nm,

 infc.name as deployment\_cluster\_nm,

 icv.description as cluster\_kubernetes\_version\_id,

 infc.nodes as cluster\_node\_no,

 infc.ram as cluster\_ram\_no,

 infc.cpu as cluster\_cpu\_no,

 infc.disk as cluster\_disk\_no,

 dp.deploy\_description as infrastructure\_desc,

 CASE

 WHEN infrq.multi\_master = 1 THEN 'Y'

 else 'N'

 END

 AS HA\_multi\_master\_flg,

 infrq.cluster\_label as cluster\_environment\_desc,

 dl.deploy\_time as deployment\_deploy\_tm,

 dl.order\_number as order\_no\_id,

 ob.order\_id as order\_id,

 dp.server as deployment\_host\_nm,

 to\_timestamp(CAST(dl."timestamp" AS double precision)) as deployment\_created\_dttm,

 dl.promotion\_stage as promotion\_stage\_cd,

 se.ship\_event\_nm as ship\_event\_nm,

 'UDANEXT' as source\_system\_cd

 FROM foreign\_udanext.deployments dp

 LEFT OUTER JOIN foreign\_udanext.deployment\_log dl ON (dl.deploy\_id = dp.id)

 LEFT OUTER JOIN foreign\_udanext.infrastructure\_tenant inft ON (inft.name=dp.tenant)

 JOIN cqm\_meta.deployment\_key dk ON (concat\_ws('|',dl.deploy\_id, dl.task\_id,'UDANEXT') = dk.deployment\_key\_txt)

 -- JOIN sel\_meta\_key dk ON concat\_ws('|',dl.deploy\_id, dl.task\_id,'UDANEXT') = dk.deployment\_key\_txt

 LEFT OUTER JOIN foreign\_udanext.infrastructure\_cluster infc ON (infc.tenant\_id=inft.id)

 LEFT OUTER JOIN foreign\_udanext.infrastructure\_cluster\_request infrq ON (infrq.id=infc.request\_id)

 LEFT OUTER JOIN foreign\_udanext.infrastructure\_cluster\_version icv ON (infc.k8s\_version\_id=icv.id)

 LEFT OUTER JOIN main.ship\_event se ON (se.ship\_event\_nm = substring(dp.shipevent from 6 for 5))

 inner join main.order\_base ob on ob.order\_no = dl.order\_number

 WHERE dk.deployment\_key\_txt NOT IN

 (SELECT source\_system\_deployment\_id

 FROM main.deployment\_base db)

2023-01-25 09:08:46 EST [354873]: user=pestor,db=cqm\_dev,app=pgAdmin 4 - CONN:5168871,client=172.16.47.57 LOG: temporary file: path "base/pgsql\_tmp/pgsql\_tmp354873.0", size 24272896

2023-01-25 09:08:46 EST [354873]: user=pestor,db=cqm\_dev,app=pgAdmin 4 - CONN:5168871,client=172.16.47.57 LOG: temporary file: path "base/pgsql\_tmp/pgsql\_tmp354873.5", size 0

2023-01-25 09:08:46 EST [354873]: user=pestor,db=cqm\_dev,app=pgAdmin 4 - CONN:5168871,client=172.16.47.57 LOG: temporary file: path "base/pgsql\_tmp/pgsql\_tmp354873.4", size 0

2023-01-25 09:08:46 EST [354873]: user=pestor,db=cqm\_dev,app=pgAdmin 4 - CONN:5168871,client=172.16.47.57 LOG: temporary file: path "base/pgsql\_tmp/pgsql\_tmp354873.3", size 0

2023-01-25 09:08:46 EST [354873]: user=pestor,db=cqm\_dev,app=pgAdmin 4 - CONN:5168871,client=172.16.47.57 LOG: temporary file: path "base/pgsql\_tmp/pgsql\_tmp354873.2", size 24231936

2023-01-25 09:08:46 EST [354873]: user=pestor,db=cqm\_dev,app=pgAdmin 4 - CONN:5168871,client=172.16.47.57 LOG: temporary file: path "base/pgsql\_tmp/pgsql\_tmp354873.1", size 24305664

2023-01-25 09:08:55 EST [354911]: user=,db=,app=,client= LOG: automatic vacuum of table "cqm\_metabase.public.qrtz\_scheduler\_state": index scans: 0

 pages: 0 removed, 1 remain, 0 skipped due to pins, 0 skipped frozen

 tuples: 56 removed, 1 remain, 0 are dead but not yet removable, oldest xmin: 53130667

 index scan not needed: 0 pages from table (0.00% of total) had 0 dead item identifiers removed

 avg read rate: 0.000 MB/s, avg write rate: 18.382 MB/s

 buffer usage: 52 hits, 0 misses, 1 dirtied

 WAL usage: 3 records, 1 full page images, 8671 bytes

 system usage: CPU: user: 0.00 s, system: 0.00 s, elapsed: 0.00 s

What is significant here is that a number of temp files are created with intermediate results. Three temp files of none-zero size are noted, each of which is slightly over 24 MB in size. It could be that, as a result of all this caching, we are running out of cache/swap space on the server which is causing the failure (albeit with an EXTREMELY misleading error message). As you can see below, the two environments are very similar in architecture except for number of CPUs.

Below are the stats for the two machines, dev and prod.

### Dev environment

$ df

Filesystem 1K-blocks Used Available Use% Mounted on

devtmpfs 32810860 0 32810860 0% /dev

tmpfs 32829716 1308 32828408 1% /dev/shm

tmpfs 32829716 1016 32828700 1% /run

tmpfs 32829716 0 32829716 0% /sys/fs/cgroup

/dev/mapper/vg\_default-lv\_root 69094696 18072504 47830792 28% /

/dev/sda1 499656 347104 115856 75% /boot

/dev/mapper/vg\_default-lv\_var 16382844 1684248 13843352 11% /var

/dev/mapper/vg\_default-lv\_tmp 5095040 23452 4793060 1% /tmp

/dev/mapper/vg\_data-lv\_data 2604006408 1295117116 1308872908 50% /var/lib/pgsql

regxbc02.unx.sas.com:/opt/sysadm 206288896 91794432 103992320 47% /opt/sysadm

isilon03nfs.unx.sas.com:/ifs/nosnaps/cqm\_backups 1468006400 1182180864 285825536 81% /var/cqm\_backups

isilon03.unx.sas.com:/ifs/data/cqmtest\_backups 1073741824 1073741824 0 100% /var/cqmtest\_backups

tmpfs 6565940 0 6565940 0% /run/user/3595

tmpfs 6565940 0 6565940 0% /run/user/10639

cat /proc/meminfo

MemTotal: 65659436 kB

MemFree: 17569556 kB

MemAvailable: 47897876 kB

Buffers: 248268 kB

Cached: 45506792 kB

SwapCached: 14000 kB

Active: 28747392 kB

Inactive: 17297648 kB

Active(anon): 16033288 kB

Inactive(anon): 266548 kB

Active(file): 12714104 kB

Inactive(file): 17031100 kB

Unevictable: 0 kB

Mlocked: 0 kB

SwapTotal: 4194300 kB

SwapFree: 3939324 kB

Dirty: 52 kB

Writeback: 0 kB

AnonPages: 280560 kB

Mapped: 16046604 kB

Shmem: 16009848 kB

KReclaimable: 1309076 kB

Slab: 1543072 kB

SReclaimable: 1309076 kB

SUnreclaim: 233996 kB

KernelStack: 6432 kB

PageTables: 225972 kB

NFS\_Unstable: 0 kB

Bounce: 0 kB

WritebackTmp: 0 kB

CommitLimit: 37024016 kB

Committed\_AS: 17046424 kB

VmallocTotal: 34359738367 kB

VmallocUsed: 0 kB

VmallocChunk: 0 kB

Percpu: 107008 kB

HardwareCorrupted: 0 kB

AnonHugePages: 159744 kB

ShmemHugePages: 0 kB

ShmemPmdMapped: 0 kB

FileHugePages: 0 kB

FilePmdMapped: 0 kB

HugePages\_Total: 0

HugePages\_Free: 0

HugePages\_Rsvd: 0

HugePages\_Surp: 0

Hugepagesize: 2048 kB

Hugetlb: 0 kB

DirectMap4k: 542528 kB

DirectMap2M: 50837504 kB

DirectMap1G: 17825792 kB

$ lscpu | egrep 'Model name|Socket|Thread|NUMA|CPU\(s\)'

CPU(s): 4

On-line CPU(s) list: 0-3

Thread(s) per core: 1

Socket(s): 4

NUMA node(s): 1

Model name: Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz

NUMA node0 CPU(s): 0-3

select version()

PostgreSQL 14.3 on x86\_64-pc-linux-gnu, compiled by gcc (GCC) 8.5.0 20210514 (Red Hat 8.5.0-10), 64-bit

### Prod environment

$ df

Filesystem 1K-blocks Used Available Use% Mounted on

devtmpfs 65840984 0 65840984 0% /dev

tmpfs 65859840 58564 65801276 1% /dev/shm

tmpfs 65859840 1132 65858708 1% /run

tmpfs 65859840 0 65859840 0% /sys/fs/cgroup

/dev/mapper/vg\_default-lv\_root 69094696 4753588 61149708 8% /

/dev/sda1 499656 347064 115896 75% /boot

/dev/mapper/vg\_default-lv\_var 16382844 3419088 12108512 23% /var

/dev/mapper/vg\_default-lv\_tmp 5095040 20492 4796020 1% /tmp

regxbc02.unx.sas.com:/opt/sysadm 206288896 91794432 103992320 47% /opt/sysadm

tmpfs 13171968 0 13171968 0% /run/user/0

isilon03nfs.unx.sas.com:/ifs/nosnaps/cqm\_backups 1468006400 1182180864 285825536 81% /var/cqm\_backups

/dev/drbd0 1878122868 901901912 976204572 49% /var/lib/pgsql

tmpfs 13171968 0 13171968 0% /run/user/3595

tmpfs 13171968 0 13171968 0% /run/user/10639

$ cat /proc/meminfo

MemTotal: 131719684 kB

MemFree: 1893884 kB

MemAvailable: 96026960 kB

Buffers: 55028 kB

Cached: 125785868 kB

SwapCached: 20956 kB

Active: 62095452 kB

Inactive: 64324096 kB

Active(anon): 32042800 kB

Inactive(anon): 565988 kB

Active(file): 30052652 kB

Inactive(file): 63758108 kB

Unevictable: 165980 kB

Mlocked: 165980 kB

SwapTotal: 4194300 kB

SwapFree: 3481264 kB

Dirty: 2056 kB

Writeback: 0 kB

AnonPages: 734624 kB

Mapped: 31501228 kB

Shmem: 32014476 kB

KReclaimable: 1574668 kB

Slab: 1860200 kB

SReclaimable: 1574668 kB

SUnreclaim: 285532 kB

KernelStack: 9168 kB

PageTables: 738928 kB

NFS\_Unstable: 0 kB

Bounce: 0 kB

WritebackTmp: 0 kB

CommitLimit: 70054140 kB

Committed\_AS: 34301304 kB

VmallocTotal: 34359738367 kB

VmallocUsed: 0 kB

VmallocChunk: 0 kB

Percpu: 158208 kB

HardwareCorrupted: 0 kB

AnonHugePages: 376832 kB

ShmemHugePages: 0 kB

ShmemPmdMapped: 0 kB

FileHugePages: 0 kB

FilePmdMapped: 0 kB

HugePages\_Total: 0

HugePages\_Free: 0

HugePages\_Rsvd: 0

HugePages\_Surp: 0

Hugepagesize: 2048 kB

Hugetlb: 0 kB

DirectMap4k: 2967360 kB

DirectMap2M: 92452864 kB

DirectMap1G: 40894464 kB

$ lscpu | egrep 'Model name|Socket|Thread|NUMA|CPU\(s\)'

CPU(s): 16

On-line CPU(s) list: 0-15

Thread(s) per core: 1

Socket(s): 16

NUMA node(s): 1

Model name: Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz

NUMA node0 CPU(s): 0-15

Select version()

PostgreSQL 14.3 on x86\_64-pc-linux-gnu, compiled by gcc (GCC) 8.5.0 20210514 (Red Hat 8.5.0-10), 64-bit