# More space-efficient data structures for detecting duplicates in IdMapper #4

Merged **tinwelint** merged 1 commit into `neo4j:2.2` from `tinwelint:2.2-less-mem-idmapper-collisions` Apr 16, 2015

Conversation 0 Commits 1 Files changed 7 +368 −106

**tinwelint** commented Apr 16, 2015 Collaborator

Big import data sets may have a large amount of collisions (accidental or actual duplicates). Detecting duplicate input ids within the same group was previously done using a combination of maps, although that could quickly run out of heap memory.

This commit introduces another way of doing this detection. Basically it works by copying the subset of collisions into a new cache (NumberArray so an live off-heap) with its own tracker cache associated with it. This pair of arrays will be sorted with ParallelSort just like the whole data set was sorted just previously. Given the now sorted tracker cache over the collisions and the kept input ids for these collisions, all potential duplicates are next to each other and can be compared in isolation.

co-author: **@alexaverbuch**

**tinwelint** added kernel 2.2 labels Apr 16, 2015

`More space-efficient data structures for detecting duplicates in IdMa…` `92f8f18`

**alexaverbuch** commented on `92f8f18` Apr 16, 2015 Collaborator

pending green build, everything looks ok to me

**tinwelint** merged commit **9ababa5** into `neo4j:2.2` Apr 16, 2015

1 check passed

View details

Sign up for free **to join this conversation on GitHub**. Already have an account? Sign in to comment

**Labels**

2.2 kernel

**Milestone**

No milestone

**Assignee**

No one assigned

**2 participants**