(Backport) Import tool has ability to skip duplicate nodes

Merged **tinwelint** merged 1 commit into `neo4j:2.2` from `tinwelint:2.2-backport-bip-skip-dup-nodes` Apr 21, 2015

Conversation 1 Commits 1 Files changed 34 +756 −244

tinwelint commented Apr 21, 2015 Collaborator

rather nodes that have the same input ids in the same group. This ability is controlled with `--skip-duplicate-nodes<=true/false>`, whereas the skipping of bad relationships `--skip-bad-relationships<=true/false>`. All and any allowed bad entities (both relationships and duplicate nodes) are collected and controlled using `--bad` and `--bad-tolerance`.

Duplicate nodes are detected after node stage, when preparing IdMapper and deleted as a side-effect in node->relationship stage. For this the handling of not-in-use records is controlled using inUse record flag as opposed to null record due to a detail in the BatchingPageCache. It's generally better to not leave a record to chance, rather write it as inUse or not.

(cherry picked from commit `a56eb15`)

Conflicts:
community/import-tool/src/main/java/org/neo4j/tooling/ImportTool.java
community/import-tool/src/test/java/org/neo4j/tooling/ImportToolDocIT.java
community/import-tool/src/test/java/org/neo4j/tooling/ImportToolTest.java
community/kernel/src/main/java/org/neo4j/unsafe/impl/batchimport/ParallelBatchImporter.java
community/kernel/src/main/java/org/neo4j/unsafe/impl/batchimport/RelationshipEncoderStep.java
community/kernel/src/main/java/org/neo4j/unsafe/impl/batchimport/cache/idmapping/string/EncodingIdMapper.java
community/kernel/src/main/java/org/neo4j/unsafe/impl/batchimport/input/Group.java
community/kernel/src/main/java/org/neo4j/unsafe/impl/batchimport/input/Inputs.java
community/kernel/src/test/java/org/neo4j/unsafe/impl/batchimport/cache/idmapping/string/EncodingIdMapperTest.java

tinwelint added **kernel** **2.2** labels Apr 21, 2015

tinwelint commented Apr 21, 2015 Collaborator

Having this commit in 2.3 *only* is incredibly annoying, conflict-wise. Any change to batch importer conflicts. So backporting this single commit that is the difference between 2.2 and 2.3. We can choose to disable its functionality in `ImportTool` in 2.2 if we want to though.

Should be forward-merged as a null-merge to 2.3.

`Import tool has ability to skip duplicate nodes` … f79caeb

tinwelint merged commit **a4516d1** into `neo4j:2.2` Apr 21, 2015 **View details**
1 check passed

tinwelint deleted the `tinwelint:2.2-backport-bip-skip-dup-nodes` branch Apr 21, 2015

Labels
2.2
kernel

Milestone
No milestone

Assignee
No one assigned

1 participant