



This repository Search

Explore Features Enterprise Blog

Sign up

Sign in

neo4j / neo4j

Watch 238

Star 1,885

Fork 730

Better ordering in generated dumps #3

Closed timmytofu wants to merge 1 commit into neo4j:2.3 from timmytofu:efficient-dumps

Conversation 14

Commits 1

Files changed 2

+7 -7



timmytofu commented Apr 15, 2015

Would this order not be better for a dump file? Nodes are imported first, without constraints or indices slowing things down, then constraints and indices are added so the relationship creation can use them (maybe, if they're ready)?

The only downside I can think of is if someone is importing into an already populated database and one of the nodes conflicts with a constraint, you'll wait and see the error after a bit of work has already been done. That seems like a rare situation, though.

Switches order of entities exported in dump - first nodes, then indic... 96a2124



timmytofu commented Apr 15, 2015

I can't see the error, but I could well have borked a test, I couldn't actually get the tests running in a reasonable amount of time.

Related neo4j#2491 neo4j#2625



jotomo commented Apr 16, 2015

@timmytofu The build's most likely failing because you're not on the CI's whitelist, so he refuses to do the build.

What kind of speedups have you measured?

lutovich added cypher 2.3 labels Apr 17, 2015



jakewins commented Apr 22, 2015

Collaborator

+1, what @jotomo said. Would you mind signing the CLA so that we're allowed to include your contributions? It's a quick email, instructions here: <http://neo4j.com/docs/stable/cla.html>

As per the contents of the PR - @jexp @boggle y'all got any feedback on this? I'm not familiar with the dump code..



jexp commented Apr 23, 2015

Collaborator

It's an interesting suggestion but that only holds true for imports in empty databases, if the database is already populated the constraints have to go in first. Otherwise you might end up with duplicates.

I wanted to rework the dump code anyway to be more sensible. Which is now possible but was not back then.



boggle commented Apr 23, 2015

Collaborator

In terms of performance, creating indices last might make more sense as the initial index population is very fast (background store scan) vs many additional index writes per created node. Would be great to

Labels

2.3

cypher

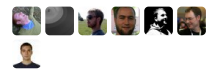
Milestone

No milestone

Assignee

No one assigned

7 participants



measure before merging this. I agree with Michael that constraints should go first to check against duplicates though.



timmytofu commented Apr 26, 2015

@jexp @boggle I mention the empty vs. populated database bit in the second half of the original message. To note, though, even putting the constraints at the top doesn't solve the problem, as if the constraint is being imported into an already populated database there might already be data in violation. Importing constraints into a populated database is inherently unpredictable, hence I still think constraints should be deferred if it does indeed increase speed - throw the weird cases under the bus to speed up the common case. Since it all happens within a transaction, neither way should result in corrupted data, one just fails more quickly by making all imports run more slowly.

On that matter I haven't been able to test it for real speed increases - I tried using our data (~1mil nodes, ~2.5mil relationships, nine distinct labels, fifteen relationship types, several indices and unique constraints) and the current dump output, but it ran for hours and hours and eventually I just gave up. We usually do our importing by copying the `data/` directory, though a reasonably performant CQL dump would be useful for porting data across versions (which we'd want to do since we can't import from CSV in 2.2.1, see #4434).

TL/DR this was just kind of a lark/brainfart. I've sent in the CLA, though, so take it for whatever it's worth.



timmytofu commented Apr 26, 2015

Actually, cla (@t) neotechnology dot cöm bounced back at me.



jakewins commented Apr 28, 2015

Collaborator

@timmytofu did you email cla@neotechnology.com, or the literal cla(@t)neotechnology.cöm? If should be the former, I think @jexp checked this earlier today, the former should work. Please let us know if it doesn't!





timmytofu commented Apr 28, 2015

I used the proper email addy:

 **timmy tofu** <timmy_tofu@linux.com>
to cla 

Hi. My name is Timothy Paul Adams (timmy_tofu@linux.com).
I agree to the terms in the attached Neo4j Contributor License Agreement.



 **Mail Delivery Subsystem** <mailer-daemon@googlemail.com>
to me 

Delivery to the following recipient failed permanently:

cla@neotechnology.com

Technical details of permanent failure:

Google tried to deliver your message, but it was rejected by the relay smtp.gmail.com by smtp.gmail.com. [74.125.22.109].

The error that the other server returned was:

535-5.7.8 Username and Password not accepted. Learn more at

535 5.7.8 <http://support.google.com/mail/bin/answer.py?answer=14257> d36sm10556409qkh.45 - gsmtpt
(SMTP AUTH failed with the remote server)

----- Original message -----

X-Received: by 10.55.31.90 with SMTP id f87mr8580434qkf.38.1430067966307; Sun,
26 Apr 2015 10:06:06 -0700 (PDT)

MIME-Version: 1.0

Received: by 10.140.41.38 with HTTP; Sun, 26 Apr 2015 10:05:45 -0700 (PDT)

From: timmy tofu <timmy_tofu@linux.com>

Date: Sun, 26 Apr 2015 13:05:45 -0400

Message-ID: <CADmHUV-SmPST-q9EmazbDq+HyXQefZwOU7P0Xf7HCw-ZQKOcA@mail.gmail.com>

Subject: CLA

To: cla@neotechnology.com

Content-Type: multipart/mixed; boundary=001a1147efc023b4570514a3a588



jakewins commented May 29, 2015

Collaborator

Sorry for dropping the ball on this, I've had GH notifications set wrong.

Ping [@jexp](#) [@simpsonjulian](#), have there been issues with the CLA email?



simpsonjulian commented May 29, 2015

Owner

[@jakewins](#) we're getting CLA email; that one looks like it failed to be relayed on it's way to us.



jakewins commented May 29, 2015

Collaborator

[@boggle](#) [@jexp](#) will we merge this in? If there's performance concerns with this approach, perhaps best to close this PR.



boggle commented May 31, 2015

Collaborator

[@timmytofu](#) [@jexp](#) [@jakewins](#) Thanks for the effort and for measuring the impact. Another reason for putting constraints first would be to prevent importing data if it would violate a later constraint and also roll back importing if the data is not in the right shape. Then again, someone might want to construct a result in order to fix up a database so that the constraint holds, in which case putting constraints last seems more desirable. It seems to me that this whole area needs more thinking on our part, perhaps the order of needs to be made configurable. Closing for now and happy to re-open if there's a strong argument why some order is always preferable.



 **boggle** closed this May 31, 2015



 **jotomo** referenced this pull request May 31, 2015

Cypher exporter / shell's dump command improvements #2625

 Open

0 of 10 tasks
complete



jotomo commented May 31, 2015

FYI: I created a meta issue for all things related to dumps; I've added this issue and linked here. Issue in question in [#2625](#)

Sign up for free

to join this conversation on GitHub. Already have an account? [Sign in to comment](#)

