

Ten. Million. Questions. Let's celebrate [all we've done together](#).

Stack Overflow is a question and answer site for professional and enthusiast programmers. It's 100% free.

Take the 2-minute tour ×

Neo4j LOAD CSV extremely slow when merging relations from a label of 3mil nodes to a lable of 20 nodes

asked 4 months ago

viewed 61 times

active 4 months ago

▲ neo4j 2.2.0 in a 3-machine cluster, with ~3 million "User" nodes and ~10 "Brand" nodes

1 When loading the ~20k rows of ":LIKES" relationship by:

★

```

USING PERIODIC COMMIT 30000
LOAD CSV WITH HEADERS FROM "file://path/to/my/file.csv" AS csvLine
MATCH (user:User {id: toInt(csvLine.userId)})
MATCH (brand:Brand {id: csvLine.brandId})
MERGE (user)-[:LIKES]->(brand)

```

It never succeeded. Sometimes it took about 6 minutes then caused master-slave swap and failed with error "Transaction has been terminated." Other times it took more than 10 minutes and caused http transaction reset. Unusual garbage collection pattern had been observed when loading. Profile showed the two "MATCH"s took very few time. So it should be the MERGE that caused the slowness and the eventual failure. Unfortunately nothing further from the profile to show what was done inside the MERGE to cause the slowness.

The node counts and relationship counts both seemed reasonable. Loading other relations that link ~3 mil nodes to another ~2 mil nodes only took a couple of minutes. So one suspicion is, the problem is due to linking too many users to too few brands at the same time? Neo4j couldnot parallel in this case? Why so many GCs were triggered?

I hoped "profile" could tell more details, such as how much time has been spent on each section, and a further breakdown of how the time has been spent on, for example, Cypher compiler, or other kernel activities? Is there a way to read internal *operation* log?

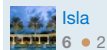
Any suggestion or ideas? Thanks!

csv neo4j load cypher

share improve this question

edited Apr 11 at 21:36

asked Apr 10 at 22:28



6 2

2 Do you have indexes created for :User(id) and :Brand(id) ? – cybersam Apr 10 at 22:37

Can you share your EXPLAIN output? (leave off periodic commit for it to work). – Michael Hunger Apr 10 at 23:37

add a comment

2 Answers

active

oldest

votes

▲ Make sure to have an index or constraint for :User(id) and :Brand(id) .

1 You can also try to replace MERGE with CREATE and assert the uniqueness before

★

```

LOAD CSV WITH HEADERS FROM "file://path/to/my/file.csv" AS csvLine
WITH distinct toInt(csvLine.userId) as userId, csvLine.brandId as brandId
MATCH (user:User {id: userId})
MATCH (brand:Brand {id: brandId})
MERGE (user)-[:LIKES]->(brand)

```

Perhaps your brand is a dense ndoe, can you try to reverse the direction of the arrow? MERGE (brand)<-[:LIKES]-(user)






Blog

 Why Stack Overflow is a Good Workplace for Women

Related

- 4 Neo4j Cypher Query Extremely Slow (about 20 minutes)
- 2 Neo4j Cypher - creating nodes and setting labels with LOAD CSV
- 2 Neo4j Cypher: MERGE conditionally with values from LOAD CSV
- 1 Create nodes and relations conditionally when loading nodes from csv in Neo4j
- 0 Loading data in Neo4j from csv
- 0 How to load ALL the columns from a *.csv into Neo4j nodes
- 1 Why would the same load csv in neo4j take exponential longer when run a second time (on a clean db)?
- 0 Labels on Nodes and Relationships from a CSV file
- 1 Neo4j Load from CSV "How to create node or relationship type dynamically as retrieved from csv?"
- 0 How to import relationships between nodes of the same label from a csv file in neo4j?

Hot Network Questions

-  Did Arnold Rimmer kill the Red Dwarf crew?
-  Scraping High-Res image tiles from the MoMA website
-  Seven overlapping circles
-  Is it 'Batman' or 'Batmans'?
-  What is aliasing and what causes it?

share improve this answer

answered Apr 10 at 23:40



Michael Hunger
25.2k ● 2 ● 21 ● 41

Yes I do have Indices. Let me try reverse linking as you suggested ... – Isla Apr 11 at 2:29

Reverse the MERGE didn't help. – Isla Apr 11 at 3:54

Michael, I could not post images for the profile or vm visualizer (don't have enough reputation yet :() ... anyway the "profile load csv" took 589682msec to finishm with the MERGE, on a 100k rows file. If replacing the MERGE with a return, it took only 12463 msec. The 100k size worked fine with other relations (even 500k was fast), but not this one. I am chunking up the csv files into multiple smaller files and trying again. Any other suggestion? – Isla Apr 11 at 4:09

add a comment



Thanks for your help Michael and cybersam. Found the reason to be the MERGE on the dense nodes. Internally neo4j 2.2.0 uses linked lists (2-way) for all nodes having the same relationship, a new MERGE means 2 linked list scan which are not indexed, so scanning on the dense node's list will take longer and longer as more relations added.

Our workaround is to use CREATE when dumping the big chunk of relations in. Then update with MERGE on a much smaller dataset periodically.

However the problem is still there and it is not always possible to CREATE. neo said they will improve it.

share improve this answer

answered Apr 16 at 17:00



Isla
6 ● 2

add a comment

- Manager sounds upset every time I inform him of a (minor) obstacle
- Can you open source firmware to closed hardware?
- Select polygons one by one and export them using arcpy
- Which Star Trek character appeared on screen with the most different ranks?
- Dullness vs. going overboard: Should I be calling people 'enfants terribles' in an academic paper?
- My 4.5 yrs old son has no dominant hand
- Important formulas in Combinatorics
- Funding government by only printing more money
- Why moving fan seems transparent?
- What does OWA stand for?
- Carrying a handgun in other countries with a US concealed carry permit
- `cond` with less redundancy
- Using Emergency Fund to Sell Upside-down Car
- Should you always minimize cognitive load
- Is Let us = Let's?
- Using siunitx, \ohm results in an italic Omega
- Weird performance increase in simple benchmark
- Compact way of writing $a + b == c$ or $a + c == b$ or $b + c == a$
- Suggested alternatives for that horrible new noun 'nice-to-have'?

Your Answer

B *I*

Sign up or [log in](#)

Sign up using Google

Sign up using Facebook

Sign up using Stack Exchange

Post as a guest

Name

Email

required, but never shown

Post Your Answer

By posting your answer, you agree to the [privacy policy](#) and [terms of service](#).

Not the answer you're looking for? Browse other questions tagged [csv](#) [neo4j](#) [load](#) [cypher](#) or [ask](#)

[tour](#) [help](#) [blog](#) [chat](#) [data](#) [legal](#) [privacy policy](#) [work here](#) [advertising info](#) [mobile](#) [contact us](#) [feedback](#)

TECHNOLOGY

Stack Overflow	Programmers	Database Administrators
Server Fault	Unix & Linux	Drupal Answers
Super User	Ask Different (Apple)	SharePoint
Web Applications	WordPress Development	User Experience
Ask Ubuntu	Geographic Information Systems	Mathematica
Webmasters	Electrical Engineering	Salesforce
Game Development	Android Enthusiasts	ExpressionEngine® Answers
TeX - LaTeX	Information Security	more (13)

LIFE / ARTS

[Photography](#)
[Science Fiction & Fantasy](#)
[Graphic Design](#)
[Movies & TV](#)
[Seasoned Advice \(cooking\)](#)
[Home Improvement](#)
[Personal Finance & Money](#)
[Academia](#)
[more \(9\)](#)

CULTURE / RECREATION

[English Language & Usage](#)
[Skeptics](#)
[Mi Yodeya \(Judaism\)](#)
[Travel](#)
[Christianity](#)
[Arqade \(gaming\)](#)
[Bicycles](#)
[Role-playing Games](#)
[more \(21\)](#)

SCIENCE

[Mathematics](#)
[Cross Validated \(stats\)](#)
[Theoretical Computer Science](#)
[Physics](#)
[MathOverflow](#)
[Chemistry](#)
[Biology](#)
[more \(5\)](#)

OTHER

[Stack Apps](#)
[Meta Stack Exchange](#)
[Area 51](#)
[Stack Overflow Careers](#)