# Metacat - Bug #238

## query performance

06/19/2001 12:10 PM - Matt Jones

| | | | |
|---|---|---|---|
| **Status:** | Resolved | **Start date:** | 06/19/2001 |
| **Priority:** | Immediate | **Due date:** | |
| **Assignee:** | Jing Tao | **% Done:** | 0% |
| **Category:** | metacat | **Estimated time:** | 0.00 hour |
| **Target version:** | 1.3 | **Spent time:** | 0.00 hour |
| **Bugzilla-Id:** | 238 | | |

**Description**

Metacat has horrible query performance, and we need to fix it. It can take as much as 8 or 9 seconds to search a few megabytes of XML. See the Internet Computing paper on metacat for performance details. Dan has shown that an in memory hash table in Java outperforms metacat by a substantial margin. So, to fix this bug we need to:

1) determine why performance is so bad in metacat
2) design and implement an alternative solution

**Related issues:**

| | | |
|---|---|---|
| Blocked by Metacat - Bug #193: evaluate recursive search performance | **Resolved** | **04/09/2001** |

### History

#### #1 - 03/21/2002 05:31 PM - Jing Tao

If user want put '%' in the search value like this way:
SELECT docid,docname,doctype,date_created, date_updated, rev FROM xml_documents WHERE docid IN ((SELECT DISTINCT docid FROM xml_nodes WHERE UPPER LIKE '%%%' ))

The query will be changed to:
SELECT docid,docname,doctype,date_created, date_updated, rev FROM xml_documents

We hope this can improve performance. The result is:
In metacat: former took 58.74 second. Latter took 53.95 seconds.

In sqlplus former took 00:01:36:43, Latter took 00:01:26.25

The reseaon for so long is there are 10022 documents returned.
There is a small improvement. I am not sure if we need to check in the file to cvs.

#### #2 - 03/21/2002 07:23 PM - Matt Jones

Added myself back as QA contact. It should not have been removed in the first place.

#### #3 - 04/15/2002 10:05 AM - Jing Tao

During playing query performance, it is found that more time will take, if the result set is bigger. So performance will bad if huge result will be got.

So if we just query subset of result at one time rather than whole result set, the performance will be improved.

There are two ways to handle the subset result:
1: Cache them into memory and send it to client when they are required. This way will invole lots of session managment.
2: Just open a outstream, keep adding the subset set result into stream and send it to client.( Dan's idea).

#### #4 - 07/22/2002 08:01 AM - Jing Tao

Just now I played with metacat.nceas.ucsb.edu/knb/index.html page. It will send a query:

SELECT docid,docname,doctype,date_created, date_updated, rev FROM
xml_documents WHERE docid IN ((SELECT DISTINCT docid FROM xml_nodes WHERE
UPPER LIKE '%%%' ))

to metacat in knb and get a doclist. The total time is 56 seconds ( measured
by my watch) to get the doc list page.

But in sqlplus, I typed the query and it took 7.91 seconds to get the
documents.

So from this point, query time is only small part of time consuming process.
We need to figure out which take most of time.

#### #5 - 09/09/2002 04:09 PM - Jing Tao

By carefully measured the time consumming in search action, it is found that
the most time consuming process is permission checking. In findDocument in
DBQuery class, there are two times permission checking, the first time is
checking for the doc gotted by running query in xml_documents. The second one
is for the docs get by running extened query. Actually, extended query is
based on the docid list which got from first query. So second checking is
redundant. We can delete it.

Here is a example:

| permission checking time | Total time(seconds) | document number in result set |
|---|---|---|
| 2 | 17 | 43 |
| 1 | 13(or 12) | 43 |
| 0 | 3 | 45 |

From this table, we can see the permission checking is the main control. In
Metacat, persmission check in base on single docid, so it is easy to explain
why it took a long time when return a big set of documents.

From the printing result, it is known that to a single docid, the longest
permission checking time is about 150 milli seconds, the shortest is 8 milli
seconds, the most common is 40 -50 milli seconds. Currently there are 2219
docids in KNB site. We can imagine it would take a long time to run % search.

#### #6 - 10/09/2002 05:24 PM - Jing Tao

Matt's idea that we don't check permission by Java code and directly select
doicd from xml_access is very good.

The search query looks like:
SELECT docid,docname,doctype,date_created, date_updated, rev FROM
xml_documents WHERE docid IN ((((SELECT DISTINCT docid FROM xml_nodes WHERE
UPPER LIKE '%%%' ))))

If we intersect the docids which user has permssion, this will be very helpful.
Here is my idea:
1, The first part that user has permission is:
onwer's docid and it can be done by select docid from xml_documents table.
2, The second part that user has permission is from xml_access table:
If a user (including group and public) has a permission to read a docid, it
should have:
A. An allow rule.
B. Doesn't have a deny rule and perm_order is "allow" first.
So second part will look like:
docid in (select docid where perm_type = 'allow' and principal_name = '...')
AND docid not in (select docid where perm_type ='deny and principal_name='...'
and perm_order ='allowFirst')

Union the two parts for permission and then intersect them to oringal query,
we can get the result.

Here is a query example:
SELECT docid,docname,doctype,date_created, date_updated, rev FROM
xml_documents WHERE docid IN ((((SELECT DISTINCT docid FROM xml_nodes WHERE
UPPER LIKE '%%%' ))))
AND
(docid IN(SELECT docid FROM xml_documents WHERE user_owner ='public' OR
user_owner ='uid=jtao,o=LTER,dc=ecoinformatics,dc=org')
OR
(docid IN (SELECT docid from xml_access
WHERE (principal_name = 'uid=jtao,o=LTER,dc=ecoinformatics,dc=org' AND

perm_type = 'allow')OR (principal_name = 'public' AND perm_type = 'allow'))
AND docid NOT IN (SELECT docid
from xml_access WHERE (principal_name =
'uid=jtao,o=LTER,dc=ecoinformatics,dc=org' AND perm_type = 'deny' AND
perm_order ='allowFirst')OR (principal_name = 'public' AND perm_type = 'deny'
AND perm_order ='allowFirst'))))

Here is result for % search query:

| Return packages | time for old check(sec) | time for new check(sec) |
|---|---|---|
| 51 | 16 | 2 |
| 51 | 15 | 3 |
| 51 | 16 | 3 |

Comments and suggestions?

#### #7 - 01/23/2003 02:34 PM - Jing Tao

In production metacat, do a percentage search it now take about 20 seconds rather than 150 seconds.

#### #8 - 03/27/2013 02:13 PM - Redmine Admin

Original Bugzilla ID was 238