BIEN 3 - Task #905

Task # 887 (Rejected): fix disk space leak that fills the disk and crashes the import

Task # 902 (Resolved): fix bug that causes joining on the wrong columns in the import

narrow down the cause of the import bug (incorrect join columns and disk space leak)

04/25/2014 05:58 AM - Aaron Marcuse-Kubitza

Status:	Resolved	**Start date:**	
Priority:	High	**Due date:**	
Assignee:	Aaron Marcuse-Kubitza	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:			
Activity type:			

Description

see #887, #902

~~alternate OS approach~~

tried, and problem also occurs on Mac, so using other approaches

1. ~~fix `make install`, which sets up the entire BIEN installation and dependencies~~
2. ~~test the import on the local testing machine (a Mac), which already has most of the dependencies~~
3. ~~if that doesn't work, try the other approaches below~~

~~working-backup approach~~

restoring to a working backup allows us to successfully run the import, so that we can then test system changes to see what broke things

1. ~~restore vegbiendev to last working configuration~~
2. ~~get restored VM to work on VirtualBox~~
    1. ~~install bootloader~~
    2. ~~install device drivers~~*the Linux VM configuration does not support the VirtualBox ethernet device natively, so it must be configured manually*
3. ~~set up database configuration[1] (the files and data that are not part of a Postgres-level backup)~~
4. ~~install database contents~~[1]
   there are 2 possible strategies for this:
    - ~~a) fix the bugs in the version of make install available in that backup, and run it to build the database from the flat files. the advantage of this approach is that make install is very fast (~15 min) when restoring only a single datasource in addition to the metadata datasources, while pg_restore would take a day or more or require a custom script to do a selective restore.~~ *not using this method*
    - ~~b) find the nearest database backup, restore it, and undo/redo modifications between the time of the VM backup and the time of the database backup. this involves either restoring the full DB (~1 day on vegbiendev, likely slower on an external USB drive) or writing a script to do a selective restore of the data needed for the test import; and then figuring out which modifications were made to the DB before/since the VM backup. (see README.TXT -> to set up the VM)~~
5. ~~test import and verify that it works~~
6. ~~perform system upgrades and see if import still works (see README.TXT -> to install system updates)~~*the import does still work[2], so the join columns bug (#902) is fortunately not in a system upgrade.*
    - ~~if not, selectively install system upgrades to narrow down the problem~~*N/A*
7. ~~update svn and see if import still works (see README.TXT -> to test a different revision)~~*it doesn't, so this means the bug is somewhere in our code*
    - ~~if not, test different revisions to narrow down the problem~~*the problem was caused by r12153, so debugging can be limited to the code affected by this commit*

[1] the backups from that time did not include the database because of incorrect information in the Postgres documentation stating that a filesystem-level backup of the live database would not be useful (in fact Postgres usually keeps around enough WAL files[3], and can be configured to keep around even more of them, so that a filesystem-level backup would work perfectly fine in most cases). this incorrect information (essentially a disclaimer) is likely there because the Postgres developers do not want users being upset at them if their filesystem-level backup happens to be unrestorable.

[2] note that there is a slowdown, where the import freezes for 25 min on a query, before resuming:

```
pg_stat_activity snapshot taken: "15:34:51"

"query_start"
"2014-07-14 14:56:59.482365-07"
"backend_start"
"2014-07-14 14:43:41.845428-07"

query running for 25 min ("0:23:56.119442" in inputs/ARIZ/omoccurrences/logs/test_import.log.sql,
but strangely 40 min ("14:56:59".."15:34:51") according to pg_stat_activity)

[2] DB query: non-cacheable:
Took 0:23:56.119442 sec
/*ARIZ.omoccurrences*/
CREATE TEMP TABLE "coordinates_pkeys" AS
SELECT
"in#41"."*occurrenceID"
, "coordinates"."coordinates_id" AS "out.coordinates_id"
FROM "in#41"
JOIN "coordinates" ON
"coordinates"."source_id" = 5
AND "coordinates"."latitude_deg" = "in#41"."_nullIf(decimalLatitude).result"
AND "coordinates"."longitude_deg" = "in#41"."_nullIf(decimalLongitude).result"
AND COALESCE("coordinates"."verbatimlatitude", CAST('\N' AS text)) = COALESCE(NULL, CAST('\N' AS
text))
AND COALESCE("coordinates"."verbatimlongitude", CAST('\N' AS text)) = COALESCE(NULL, CAST('\N' AS
text))
AND COALESCE("coordinates"."verbatimcoordinates", CAST('\N' AS text)) = COALESCE("in#41".
"ARIZ.omoccurrences.verbatimCoordinates", CAST('\N' AS text))
AND COALESCE("coordinates"."footprintgeometry_dwc", CAST('\N' AS text)) = COALESCE("in#41".
"ARIZ.omoccurrences.footprintWKT", CAST('\N' AS text))
/* EXPLAIN:
Nested Loop  (cost=5333.13..5366.19 rows=1 width=8)
  ->  Index Scan using coordinates_unique on coordinates  (cost=0.41..18.01 rows=1 width=84)
        Index Cond: ((source_id = 5) AND (COALESCE(verbatimlatitude, '\N'::text) = '\N'::text) AND
 (COALESCE(verbatimlongitude, '\N'::text) = '\N'::text))
  ->  Bitmap Heap Scan on "in#41"  (cost=5332.72..5348.17 rows=1 width=73)
        Recheck Cond: ((COALESCE("ARIZ.omoccurrences.verbatimCoordinates", '\N'::text) = COALESCE(
coordinates.verbatimcoordinates, '\N'::text)) AND (COALESCE("ARIZ.omoccurrences.footprintWKT", '\N
'::text) = COALESCE(coordinates.footprintgeometry_dwc, '\N'::text)))
        Filter: ((coordinates.latitude_deg = "_nullIf(decimalLatitude).result") AND (coordinates.l
ongitude_deg = "_nullIf(decimalLongitude).result"))
        ->  Bitmap Index Scan on "in#41_coalesce_coalesce1_coalesce2_coalesce3_idx"  (cost=0.00..5
332.72 rows=4 width=0)
              Index Cond: ((COALESCE("ARIZ.omoccurrences.verbatimCoordinates", '\N'::text) = COALE
SCE(coordinates.verbatimcoordinates, '\N'::text)) AND (COALESCE("ARIZ.omoccurrences.footprintWKT",
 '\N'::text) = COALESCE(coordinates.footprintgeometry_dwc, '\N'::text)))
*/
```

[3] except when the database is under very heavy load, such as at the beginning of full-database import

tracing approach

1. attempt to trace where the incorrect columns list is being introduced
    - generally time-consuming to trace data through complex code
    - narrowing down the affected section of code (using the working-backup approach) would reduce the amount of code that needs to be traced

rewriting approach

1. rewrite the section of code with the bug
    - in this case, that code section is the most complex part of our codebase, so this would be potentially very time-consuming
    - narrowing down the affected section of code (using the working-backup approach) would reduce the amount of code that needs to be rewritten

---

~~clean VM approach~~

not using this approach because we have a backup from the time of the last successful import, so we don't need to obtain/install dependencies in the versions they had in the last successful import

1. ~~prepare clean VMs~~
2. ~~fix `make install`, which sets up the entire BIEN installation and dependencies~~
    - normally, we do not reinstall the DB from scratch, so the bugs in `make install` only become apparent when it is run on a partial installation
3. install the database from scratch on a clean VM (VirtualBox)
    - this would involve adding any missing dependencies to our install scripts
4. test the import in the clean VM with a sample datasource to see if that reproduces the problem
    - if it does, we know it's a bug in Postgres/Ubuntu and can troubleshoot using VM images with different Postgres/Ubuntu versions
    - if it doesn't, it's a problem specific to just the vegbiendev VM and we would reset the vegbiendev VM to a clean Ubuntu install and reinstall our dependencies

~~Postgres rollback approach~~

not using this approach because it only works if the problem is with Postgres, but it could also be a problem in the other import code (eg. Python)

1. roll back Postgres to the version it was at in the last successful import
    - this may require building Postgres from source, because past *revisions* of the same numeric version might only be available in version control, not in binary form via apt-get (which numbers packages by numeric version)
    - if this isn't possible, it may be necessary to downgrade to Postgres 9.2 (which will unfortunately be missing some features that we now use)
2. see if this fixes the problem

History

#1 - 04/30/2014 02:21 PM - Aaron Marcuse-Kubitza

- Description updated

#2 - 04/30/2014 02:21 PM - Aaron Marcuse-Kubitza

- Description updated

#3 - 05/03/2014 01:23 AM - Aaron Marcuse-Kubitza

- Description updated

- % Done changed from 0 to 30

#4 - 05/09/2014 07:36 PM - Aaron Marcuse-Kubitza

- Description updated

#5 - 05/09/2014 07:37 PM - Aaron Marcuse-Kubitza

- % Done changed from 30 to 40

#6 - 05/28/2014 02:57 PM - Aaron Marcuse-Kubitza

- Description updated

#7 - 05/28/2014 02:59 PM - Aaron Marcuse-Kubitza

- Description updated

#8 - 05/28/2014 03:01 PM - Aaron Marcuse-Kubitza

- Parent task set to #902

#9 - 05/28/2014 03:01 PM - Aaron Marcuse-Kubitza

- % Done changed from 20 to 40

#10 - 05/28/2014 03:06 PM - Aaron Marcuse-Kubitza

- Description updated

#11 - 05/28/2014 03:55 PM - Aaron Marcuse-Kubitza

- Description updated

- % Done changed from 40 to 50

#12 - 07/11/2014 10:48 AM - Aaron Marcuse-Kubitza

- Description updated

#13 - 07/11/2014 12:25 PM - Aaron Marcuse-Kubitza

- Description updated

#14 - 07/11/2014 12:28 PM - Aaron Marcuse-Kubitza

- Description updated

#15 - 07/11/2014 12:30 PM - Aaron Marcuse-Kubitza

- Description updated

#16 - 07/11/2014 12:38 PM - Aaron Marcuse-Kubitza

- Description updated

#17 - 07/11/2014 12:39 PM - Aaron Marcuse-Kubitza

- Description updated

#18 - 07/11/2014 12:39 PM - Aaron Marcuse-Kubitza

- Description updated

#19 - 07/11/2014 12:40 PM - Aaron Marcuse-Kubitza

- Description updated

#20 - 07/11/2014 12:47 PM - Aaron Marcuse-Kubitza

- Description updated

#21 - 07/11/2014 12:51 PM - Aaron Marcuse-Kubitza

- Description updated

#22 - 07/11/2014 12:53 PM - Aaron Marcuse-Kubitza

- Description updated

#23 - 07/11/2014 01:02 PM - Aaron Marcuse-Kubitza

- Description updated

#24 - 07/11/2014 01:10 PM - Aaron Marcuse-Kubitza

- Description updated

#25 - 07/11/2014 01:11 PM - Aaron Marcuse-Kubitza

- Description updated

#26 - 07/11/2014 01:15 PM - Aaron Marcuse-Kubitza

- Description updated

#27 - 07/11/2014 01:26 PM - Aaron Marcuse-Kubitza

- Description updated

#28 - 07/11/2014 01:31 PM - Aaron Marcuse-Kubitza

- Description updated

#29 - 07/11/2014 01:34 PM - Aaron Marcuse-Kubitza

- Description updated

#30 - 07/11/2014 01:46 PM - Aaron Marcuse-Kubitza

- Description updated

#31 - 07/11/2014 01:47 PM - Aaron Marcuse-Kubitza

- Description updated

#32 - 07/11/2014 01:55 PM - Aaron Marcuse-Kubitza

- Description updated

#33 - 07/11/2014 01:56 PM - Aaron Marcuse-Kubitza

- Description updated

#34 - 07/11/2014 03:18 PM - Aaron Marcuse-Kubitza

- Description updated

#35 - 07/14/2014 01:28 PM - Aaron Marcuse-Kubitza

- Description updated

- % Done changed from 100 to 70

#36 - 07/14/2014 02:55 PM - Aaron Marcuse-Kubitza

- Description updated

#37 - 07/14/2014 03:17 PM - Aaron Marcuse-Kubitza

- Description updated

- % Done changed from 70 to 80

#38 - 07/14/2014 03:18 PM - Aaron Marcuse-Kubitza

- Description updated

#39 - 07/14/2014 03:37 PM - Aaron Marcuse-Kubitza

- Description updated

#40 - 07/14/2014 03:40 PM - Aaron Marcuse-Kubitza

- Description updated

#41 - 07/14/2014 04:04 PM - Aaron Marcuse-Kubitza

- Description updated

#42 - 07/14/2014 10:18 PM - Aaron Marcuse-Kubitza

- Description updated

#43 - 07/14/2014 10:18 PM - Aaron Marcuse-Kubitza

- % Done changed from 80 to 90

#44 - 07/14/2014 10:52 PM - Aaron Marcuse-Kubitza

- Subject changed from narrow down the cause of the incorrect join columns and disk space leak to narrow down the cause of the import bug (incorrect join columns and disk space leak)

#45 - 07/14/2014 11:15 PM - Aaron Marcuse-Kubitza

- Description updated

#46 - 07/14/2014 11:21 PM - Aaron Marcuse-Kubitza

- Description updated

#47 - 07/14/2014 11:34 PM - Aaron Marcuse-Kubitza

- Description updated

#48 - 07/15/2014 01:45 PM - Aaron Marcuse-Kubitza

- Description updated

#49 - 07/15/2014 05:47 PM - Aaron Marcuse-Kubitza

- Description updated

- Status changed from New to Resolved

- % Done changed from 90 to 100

bug fixed in r14074

#50 - 07/15/2014 09:51 PM - Aaron Marcuse-Kubitza

- Description updated

#51 - 07/15/2014 09:52 PM - Aaron Marcuse-Kubitza

- Description updated

#52 - 07/15/2014 09:54 PM - Aaron Marcuse-Kubitza

- Description updated

#53 - 07/15/2014 09:54 PM - Aaron Marcuse-Kubitza

- Description updated

#54 - 04/28/2017 04:39 AM - Aaron Marcuse-Kubitza

- Description updated