QGIS Application - Bug report #20033
Column misalignment after operations on large geopackage

2018-10-05 07:57 PM - belg4mit -

Status:	Closed		
Priority:	High		
Assignee:			
Category:	Processing/Core		
Affected QGIS version:	3.2.3	**Regression?:**	No
Operating System:	Windows 10	**Easy fix?:**	No
Pull Request or Patch supplied:	No	**Resolution:**	fixed/implemented
Crashes QGIS or corrupts data:	Yes	**Copied to github as #:**	27855

Description

I have been working with a large public data set from NY http://gis.ny.gov/gisdata/inventories/details.cfm?DSID=1300
Specifically, the parcel data http://gis.ny.gov/gisdata/fileserver/?DSID=1300&file=NYS-Tax-Parcel-Centroid-Points.gdb.zip
This data set contains far more columns than I need, and is i a proprietary format, so I've been converting it with ogr2ogr like so
@@ogr2ogr -f GPKG foo.gpkg -select "COUNTY_NAME, PARCEL_ADDR, CITYTOWN_NAME, LOC_UNIT, LOC_ZIP, PROP_CLASS, LAND_AV, TOTAL_AV, FULL_MARKET_VAL, YR_BLT, SEWER_DESC, WATER_DESC, UTILITIES_DESC, BLDG_STYLE_DESC, HEAT_TYPE_DESC, FUEL_TYPE_DESC, SQFT_LIVING, NBR_KITCHENS, NBR_BEDROOMS, MAIL_ADDR, PO_BOX, MAIL_CITY, MAIL_STATE, MAIL_ZIP, ADD_MAIL_ADDR, ADD_MAIL_PO_BOX, ADD_MAIL_CITY, ADD_MAIL_STATE, ADD_MAIL_ZIP, SWIS_SBL_ID, OWNER_TYPE, DUP_GEO" -where "PROP_CLASS IN" NYS_Tax_Parcel_Centroid_Points.gdb NYS_Tax_Parcel_Centroid_Points

This works fine however, unfortunately very few records include a ZIP Code, therefore I have been trying to do a join by location against Census Bureau ZCTA files. I downloaded the shape files from here https://www.census.gov/geo/maps-data/data/cbf/cbf_zcta.html, then converted them to a geopackage and removed all columns except for ZCTA5CE10 and tried to spatial join this layer with the parcel layer above. Whenever I do this (be it to the whole file, or a queries subset such as "CITYTOWN_NAME"='Ossining') the columns get mangled. Typically the first few columns are unperturbed, but then everything after is shifted to the left one, and sometimes the last column (SWIS_SBL_ID) has its values placed near the transition from normalcy.

History

#1 - 2018-10-05 10:13 PM - belg4mit -

- Affected QGIS version changed from 3.2.2 to 3.2.3

Broken in 3.2.3 as well as 3.2.2

#2 - 2018-10-06 12:22 PM - Giovanni Manghi

- Category changed from Vectors to Processing/Core

- Status changed from Open to Feedback

- Priority changed from Normal to High

I tried on master/linux, and cannot confirm. I prepared the data as you describe. Then for both inputs I saved (in gpkg) subsets to make the test run in a reasaonable amount of time. Both subsets were given the CRS 26918. The results of "join by location" using the "intersects" predicate seems ok to me.

I suggest you to prepare a project with the data and link it here, then also show us what exact parameters are you using for the join by location operation.

#3 - 2018-10-06 08:08 PM - belg4mit -

- File Borked.PNG added
- File Options.PNG added

Example files and results are available here
https://nmrgroupinc-my.sharepoint.com/:u:/r/personal/jpierce_nmrgroupinc_com/Documents/Public/20033.zip?csf=1&e=XG2fDM (22MB)

I'm using a few more predicates, but I get the same results even with just Intersects. Also, as another data point, this has happened with several different files.

#4 - 2018-10-08 10:14 AM - Giovanni Manghi

belg4mit - wrote:

> *Example files and results are available here*
> *https://nmrgroupinc-my.sharepoint.com/:u:/r/personal/jpierce_nmrgroupinc_com/Documents/Public/20033.zip?csf=1&e=XG2fDM (22MB)*

it asks me to login in a Microsoft site, which I would rather not do. Do you have another link?

#5 - 2018-10-08 03:04 PM - belg4mit -

The sharing permissions were incorrect for some reason, try this:
https://nmrgroupinc-my.sharepoint.com/:u:/p/jpierce/ETsAu_igD9xFvPrWYG6myj8BG3S_GQ2UV-dApVZsoMOEjA?e=1YvA4P

#6 - 2018-10-09 10:56 AM - Giovanni Manghi

belg4mit - wrote:

> *The sharing permissions were incorrect for some reason, try this:*
> *https://nmrgroupinc-my.sharepoint.com/:u:/p/jpierce/ETsAu_igD9xFvPrWYG6myj8BG3S_GQ2UV-dApVZsoMOEjA?e=1YvA4P*

everything looks normal here (also in your attached result), unless of course I'm not understanding what is wrong in results. But unless I'm blind (possibly) I'm no seeing it.

#7 - 2018-10-09 02:58 PM - belg4mit -

Did you check the screenshots I included? I neglected to spell it out,
but some of the fields are pretty obviously misaligned. For example, "NY"
(New York state) as the the ZIP Code, and the city of Ossining as the state.Similarly, there is no way anything has been constructed in (YR_BLT) 367400.

#8 - 2018-10-09 08:48 PM - Giovanni Manghi

belg4mit - wrote:

> *Did you check the screenshots I included? I neglected to spell it out,*
> *but some of the fields are pretty obviously misaligned. For example, "NY"*

> *(New York state) as the the ZIP Code, and the city of Ossining as the state.Similarly, there is no way anything has been constructed in (YR_BLT) 367400.*

I see that the wrong data (i.e. "NY" in the "mail_zip" field and large numbers in the "yr_blt" field) already in your INPUT layer (Ossining), so of course it yelds "wrong" results in the output of the join.

#9 - 2018-10-09 09:04 PM - belg4mit -

Hrmph, so they are. I could have sworn I checked after subsetting the original ogr2ogr output and it was fine. It seems any operation that generates a new file from these large geopackages, not just spatial join, causes the problem.

Here's the ogr2ogr output https://nmrgroupinc-my.sharepoint.com/:u:/p/jpierce/EcV_o8PsWalGgh-nuDKDRJYBd9ADzEHOHsiquZx8fkuQHg?e=3O8ytz which is properly aligned; I tried to use the Ossining subset to reduce the download and opening time (it's 810MB and almost three million records), but it has helped further frame the problem at least.

#10 - 2018-10-10 12:32 PM - Giovanni Manghi

> *Here's the ogr2ogr output https://nmrgroupinc-my.sharepoint.com/:u:/p/jpierce/EcV_o8PsWalGgh-nuDKDRJYBd9ADzEHOHsiquZx8fkuQHg?e=3O8ytz which is properly aligned; I tried to use the Ossining subset to reduce the download and opening time (it's 810MB and almost three million records), but it has helped further frame the problem at least.*

I downloaded this dataset, loaded it along with the administrative boundaries. On map I made a selection on both layers to define subsets, then in the "join by location" tool I enabled for both inputs the "selected features only" option. Used the spatial predicates you shown in the screenshot you attached here: results are correct.

#11 - 2018-10-10 01:06 PM - Giovanni Manghi

> *Here's the ogr2ogr output https://nmrgroupinc-my.sharepoint.com/:u:/p/jpierce/EcV_o8PsWalGgh-nuDKDRJYBd9ADzEHOHsiquZx8fkuQHg?e=3O8ytz which is properly aligned; I tried to use the Ossining subset to reduce the download and opening time (it's 810MB and almost three million records), but it has helped further frame the problem at least.*

redid the test with the full datasets (1846.00 seconds), results are correct here.

#12 - 2018-10-10 04:52 PM - belg4mit -

- Subject changed from Column misalignment after vector operations on large table to Column misalignment after operations on large geopackage
- File SubsetFail.PNG added
- File Attributes.PNG added

I just tried with subsetting, and it still fails here. Are you also using Windows? And generating a geopackage? Creating a geopackage from the two geopackages works, at least for a selected subset, when outputting to a shapefile.

#13 - 2018-10-10 05:14 PM - Giovanni Manghi

belg4mit - wrote:

> *I just tried with subsetting, and it still fails here. Are you also using Windows?*

I'm on Linux

> *And generating a geopackage? Creating a geopackage from the two geopackages works, at least for a selected subset, when outputting to a shapefile.*

I used as inputs this https://nmrgroupinc-my.sharepoint.com/:u:/p/jpierce/EcV_o8PsWalGgh-nuDKDRJYBd9ADzEHOHsiquZx8fkuQHg?e=3O8ytz and the boundaries you posted in a previous comment. The subset was done by selection (so I have not used QGIS or ogr2ogr to create physically a new sub-dataset).

Please test on QGIS master (I'm using QGIS master).

#14 - 2018-10-10 09:17 PM - belg4mit -

Hmm, after I was able to track down the installer, the process seems to work fine in the weekly build.

#15 - 2018-10-11 09:57 AM - Giovanni Manghi

- Status changed from Feedback to Closed

- Resolution set to fixed/implemented

Files

Options.PNG	39 KB	2018-10-06	belg4mit -
Borked.PNG	332 KB	2018-10-06	belg4mit -
SubsetFail.PNG	267 KB	2018-10-10	belg4mit -
Attributes.PNG	38.5 KB	2018-10-10	belg4mit -