Data recovery using pg_filedump

Aleksander Alekseev

```
git clone git://git.postgresql.org/git/pg_filedump.git

cd pg_filedump

make
```

./pg_filedump --help

Usage: pg_filedump [-abcdfhikxy] [-R startblock [endblock]] [-D attrlist]

[-S blocksize] [-s segsize] [-n segnumber] file

Display formatted contents of a PostgreSQL heap/index/control file

Recovering data

```
create table tt (x int, y bool, z text, w timestamp);

insert into tt values(123, true, 'Text test test', now());

insert into tt values(456, null, 'Ололо трооло', null);

checkpoint;


select relfilenode from pg_class where relname = 'tt';

-- 16393
```

./pg_filedump -D int,bool,text,timestamp /path/to/db/base/16384/16393

```
Block    0 ********************************************************
<Header> -----
 Block Offset: 0x00000000         Offsets: Lower      32 (0x0020)
 Block: Size 8192  Version    4            Upper    8080 (0x1f90)
 LSN:  logid      0 recoff 0x0301e4c0      Special  8192 (0x2000)
 Items:    2                      Free Space: 8048
 Checksum: 0x0000  Prune XID: 0x00000000  Flags: 0x0000 ()
 Length (including item array): 32
[...]
```

```
[...]

<Data> ------

 Item   1 -- Length:   56  Offset: 8136 (0x1fc8)  Flags: NORMAL

COPY: 123   t   Text test test  2017-01-17 16:25:03.448488

 Item   2 -- Length:   52  Offset: 8080 (0x1f90)  Flags: NORMAL

COPY: 456   \N  Ололо трооло    \N
```

```
pg_fiedump -D ...как..раньше... | grep COPY | \
  perl -lne 's/^COPY: //g; print;' > /tmp/copy.txt
cat /tmp/copy.txt
```

```
123 t   Text test test  2017-01-17 16:25:03.448488
456 \N  Ололо трооло    \N
```

```
create table tt2 (x int, y bool, z text, w timestamp);

copy tt2 from '/tmp/copy.txt';

select * from tt2;
```

```
  x   | y |        z         |             w
------+---+------------------+---------------------------
 123  | t | Text test test   | 2017-01-17 16:25:03.448488
 456  |   | Ололо трооло     |
(2 rows)
```

Recovering schema

src/include/catalog/pg_class.h:

```
#define RelationRelationId 1259

#define RelationRelation_Rowtype_Id  83


CATALOG(pg_class,1259) BKI_BOOTSTRAP BKI_ROWTYPE_OID(83) ...

{

    NameData    relname;        /* class name */

    Oid         relnamespace;   /* OID of namespace containing this...
```

https://www.postgresql.org/docs/9.6/static/catalog-pg-class.html

Table 50-11. pg_class Columns

Name	Type	References	Description
oid	oid		Row identifier (
relname	name		Name of the tal
relnamespace	oid	pg_namespace.oid	The OID of the
reltype	oid	pg_type.oid	The OID of the
reloftype	oid	pg_type.oid	For typed table
relowner	oid	pg_authid.oid	Owner of the re
relam	oid	pg_am.oid	If this is an inde
relfilenode	oid		Name of the on state
reltablespace	oid	pg_tablespace.oid	The tablespace relation has no

```
./pg_filedump -D name,oid,oid,oid,oid,oid,oid,~ \
  /path/to/base/16384/1259 | grep COPY | grep test
```

```
COPY: test  2200    16387   0   10  0   16385 -- relfilenode!

COPY: test  2200    16387   0   10  0   16385

COPY: test_pkey 2200    0   0   10  403 16391
```

src/include/catalog/pg_attribute.h:

```
#define AttributeRelationId  1249

#define AttributeRelation_Rowtype_Id  75


CATALOG(pg_attribute,1249) BKI_BOOTSTRAP BKI_WITHOUT_OIDS ...

{

    Oid         attrelid;       /* OID of relation containing this...
```

https://www.postgresql.org/docs/9.6/static/catalog-pg-attribute.html

Table 50-7. pg_attribute Columns

Name	Type	References	Description
attrelid	oid	pg_class.oid	The table this colu
attname	name		The column name
atttypid	oid	pg_type.oid	The data type of th
attstattarget	int4		attstattarget co should be collectec dependent. For sca of histogram bins t
attlen	int2		A copy of pg_type
attnum	int2		The number of the
attndims	int4		Number of dimensi any nonzero value
attcacheoff	int4		Always -1 in storag the row

```
./pg_filedump -D oid,name,oid,int,smallint,~ /path/to/base/16384/1249 | \
  grep COPY | grep 16385
```

```
COPY: 16385 k    23   -1   4
COPY: 16385 v    25   -1   -1
COPY: 16385 ctid     27   0    6
COPY: 16385 xmin     28   0    4
COPY: 16385 cmin     29   0    4
COPY: 16385 xmax     28   0    4
COPY: 16385 cmax     29   0    4
COPY: 16385 tableoid     26   0    4
```

23 and 25 are atttypid's. relfilenode for pg_type is 1247, see pg_type.h

```
./pg_filedump -i -D name,~ /path/to/base/16384/1247 | \
  grep -A5 -E 'OID: (23|25)'

  XMIN: 1  XMAX: 0  CID|XVAC: 0  OID: 23
  Block Id: 0  linp Index: 8   Attributes: 30   Size: 32
  infomask: 0x0909 (HASNULL|HASOID|XMIN_COMMITTED|XMAX_INVALID)
  t_bits: [0]: 0xff [1]: 0xff [2]: 0xff [3]: 0x07

COPY: int4
[...]
```

```
[...]

  XMIN: 1  XMAX: 0  CID|XVAC: 0  OID: 25

  Block Id: 0  linp Index: 10   Attributes: 30   Size: 32

  infomask: 0x0909 (HASNULL|HASOID|XMIN_COMMITTED|XMAX_INVALID)

  t_bits: [0]: 0xff [1]: 0xff [2]: 0xff [3]: 0x07


COPY: text
```

Result:

- relfilenode = 16385
- There are two columns:
  - k with type int4
  - v with type text

Fun facts:

- timetz is larger than timestamptz (int64 + int32 vs int64)
- up to 8 NULLable columns are for free
- table could be compressed by reordering columns

More info:

- **“Corruption War Stories”** by Christophe Pettus **http://www.pgcon.org/2017/schedule/events/1048.en.html**
- **“In-core Compression”** by Anastasia Lubennikova, Aleksander Alekseev **https://afiskon.github.io/pgconf2017-talk.html**
- **pg_filedump source code, decode.c file** **https://git.postgresql.org/gitweb/?p=pg_filedump.git;a=blob;f=decode.c**
- **“Database System Implementation”** by Hector Garcia-Molina, Jeffrey Ullman, Jennifer Widom **https://www.amazon.com/dp/0130402648/**
- **“Hacking PostgreSQL”** by Anastasia Lubennikova **http://postgres-edu.blogspot.ru/search/label/Hacking%20PostgreSQL**
- **“Becoming a PostgreSQL Contributor”** by Aleksander Alekseev **https://habr.ru/p/308442/**

Thank you for your attention!

https://twitter.com/afiskon
https://postgrespro.com/