Import a bigger dataset into postgreSQL

Are you bored with our 22-row `zoo` dataset? Me too! It's time to go a bit bigger! Let's import the 7.000.000+ row air-delays dataset we have used beforein the bash tutorials! Follow these steps to have it imported into your SQL database!

UPDATE: I've realized that this might be a bit complex – not difficult, just complex – so I put all the instructions into a short video too! Click here or read below:

Note: we will work in bash. If you haven't done my bash tutorial series yet, I highly recommend doing at least the first episode, but if you don't want to, it's also okay to simply follow my lead step by step below.

1. Open Terminal and login (`ssh`) to your data server!
2. Download the `flight_delays` data!

   ~~`wget http://stat-computing.org/dataexpo/2009/2007.csv.bz2`~~

   ~~*Note: If you have this already, skip forward to 5.*~~

   As of 29 December 2019, I realized that the dataset has been removed from its original place. Now, you can download the dataset using this code instead:

   `wget 46.101.230.157/sql_tutorial/2007.csv.bz2`
3. Set up `dtrx` ! That's a command line tool for unzipping stuff!

   (Note: this might have been already set up; if so, skip this step!)

   `sudo apt-get install dtrx`
4. Unzip the .csv file!

   `dtrx 2007.csv.bz2`

   Note: It will take around ~60 seconds to process the whole file, so don't worry, your Terminal is not freezing, it just needs some time.
5. Format your data!

   `cat 2007.csv |cut -d',' -f1,2,3,4,5,7,10,11,14,15,16,17,18,19 | grep -v ',NA' > sql_ready.csv`
6. Now we have to give permission to our postgreSQL user to create tables and load data into them. This will need multiple steps. Here's a gif first *(note: my username is `dataguy` – yours might be something else).*

First `sudo` to the user called `postgres` :

`sudo -u postgres -i`

Then start postgreSQL:

`psql`

The prompt will change to this:

`postgres=#` !

Type:

`ALTER USER [your_user_name] WITH SUPERUSER;`

This turns your original user into a super user.

Exiting from postgreSQL. Type:

```
\q
```

Then exit from the user called `postgres` :

```
exit
```

Finally access your original user's postgreSQL database from the command line:

```
psql -d postgres
```

Okay, this was the hard part…

7. Now all you need to do is create the table by simply copy-pasting these lines into your terminal:

```
CREATE TABLE flight_delays (
  year INTEGER,
  month INTEGER,
  dayofmonth INTEGER,
  dayofweek INTEGER,
  deptime INTEGER,
  arrtime INTEGER,
  flightnum INTEGER,
  tailnum VARCHAR,
  airtime INTEGER,
  arrdelay INTEGER,
  depdelay INTEGER,
  origin VARCHAR,
  dest VARCHAR,
  distance INTEGER);
```

8. And finally, copy the data from the .csv file you have just downloaded!

```
COPY flight_delays FROM '/home/tomi/sql_ready.csv' DELIMITER ',' CSV HEADER;
```

Note: make sure that you type your user name where I've typed `tomi` *or* `dataguy` *(which are my user names…)*

9. Go back to SQL Workbench and make a simple `SELECT` statement… but make sure that you use the `LIMIT` clause, too. Why?
   Because now you have over 7.000.000 rows of data. PostgreSQL can handle it easily, sure, but your computer might be frozen if you try print all that data on your screen.

   So try something like this first:

```
SELECT * FROM flight_delays LIMIT 10;
```

Statement 1 | Statement 2 | Database Explorer 3

```
1 SELECT * FROM test_test LIMIT 10;
```

Result 1 | Messages

year	month	dayofmonth	dayofweek	deptime	arrtime	flightnum	tailnum	airtime	arrdelay	depdelay	origin	dest	distance
2007	4	3	2	2108	2252	2947	89489E	68	-3	-2	DTW	HPN	505
2007	4	4	3	2108	2309	2947	80299E	74	14	-2	DTW	HPN	505
2007	4	5	4	2106	2242	2947	84239E	75	-13	-4	DTW	HPN	505
2007	4	6	5	2106	2355	2947	80339E	82	60	-4	DTW	HPN	505
2007	4	8	7	2107	2308	2947	84329E	76	13	-3	DTW	HPN	505
2007	4	9	1	2107	2247	2947	80269E	79	-9	-3	DTW	HPN	505
2007	4	10	2	2105	2241	2947	87189E	73	-15	-5	DTW	HPN	505
2007	4	11	3	2105	2245	2947	87909E	74	-11	-5	DTW	HPN	505
2007	4	12	4	2105	2250	2947	80019E	73	-6	-5	DTW	HPN	505
2007	4	13	5	2106	2253	2947	89699E	69	-3	-4	DTW	HPN	505