PostgreSQL

Uporaba IT za raziskovanje baz podatkov

Matej Kovačič
matej.kovacic@ijs.si

Jožef Stefan Institute
Centre for Knowledge Transfer
in Information Technologies
Artificial Intelligence Laboratory

SQL in PostgreSQL

SQL (angl. *Structured Query Language*) – strukturirani povpraševalni jezik za delo s podatkovnimi bazami.

PostgreSQL je odprtokodna aplikacija za upravljanje s podatkovnimi bazami, ki se v zadnjih različicah lahko kosa tudi z najbolj zmogljivimi komercialnimi rešitvami.

Lastnosti:

- zanesljivost delovanja,
- hitrost,
- zmožnost obdelave zelo velikih količin podatkov (govorimo o več milijonih zapisov),
- z dodatki je mogoče obdelovati tudi prostorske podatke (PostGIS), slike (PostPic), indeksirati besedila (OpenFTS), itd.
- brezplačna.

Organizacija podatkov

Baze → tabele → stolpci → vrstice.

```
finance=# \d
List of relations
Schema | Name     | Type  | Owner
-------+----------+-------+-------
public | placila  | table | root
public | podjetja | table | root
public | trr      | table | root
(3 rows)
```

```
finance=# \d placila
Table "public.placila"
Column        | Type | Modifiers
--------------+------+-----------
imetnik       | text |
prejemnik     | text |
prejemnik_trr | text |
znesek        | text |
valuta        | text |
datum         | text |
namen         | text |
```

Glavni podatkovni tipi

Najpogosteje uporabljeni za namene novinarskih podatkovnih analiz:

- tekstovni podatkovni tip (*text*, *char*...)
- številski podatkovni tip (*decimal*, *numeric*, *integer*,...)
- časovni podatkovni tip (*timestamp* (z ali brez časovne cone), *date*, *time*...),
- ostali (*geometrični*, *logični*,...).

Namestitev

PostgreSQL je na voljo za številne operacijske sisteme (Linux, Windows, Mac OS, BSD, Solaris).

<https://www.postgresql.org/download>

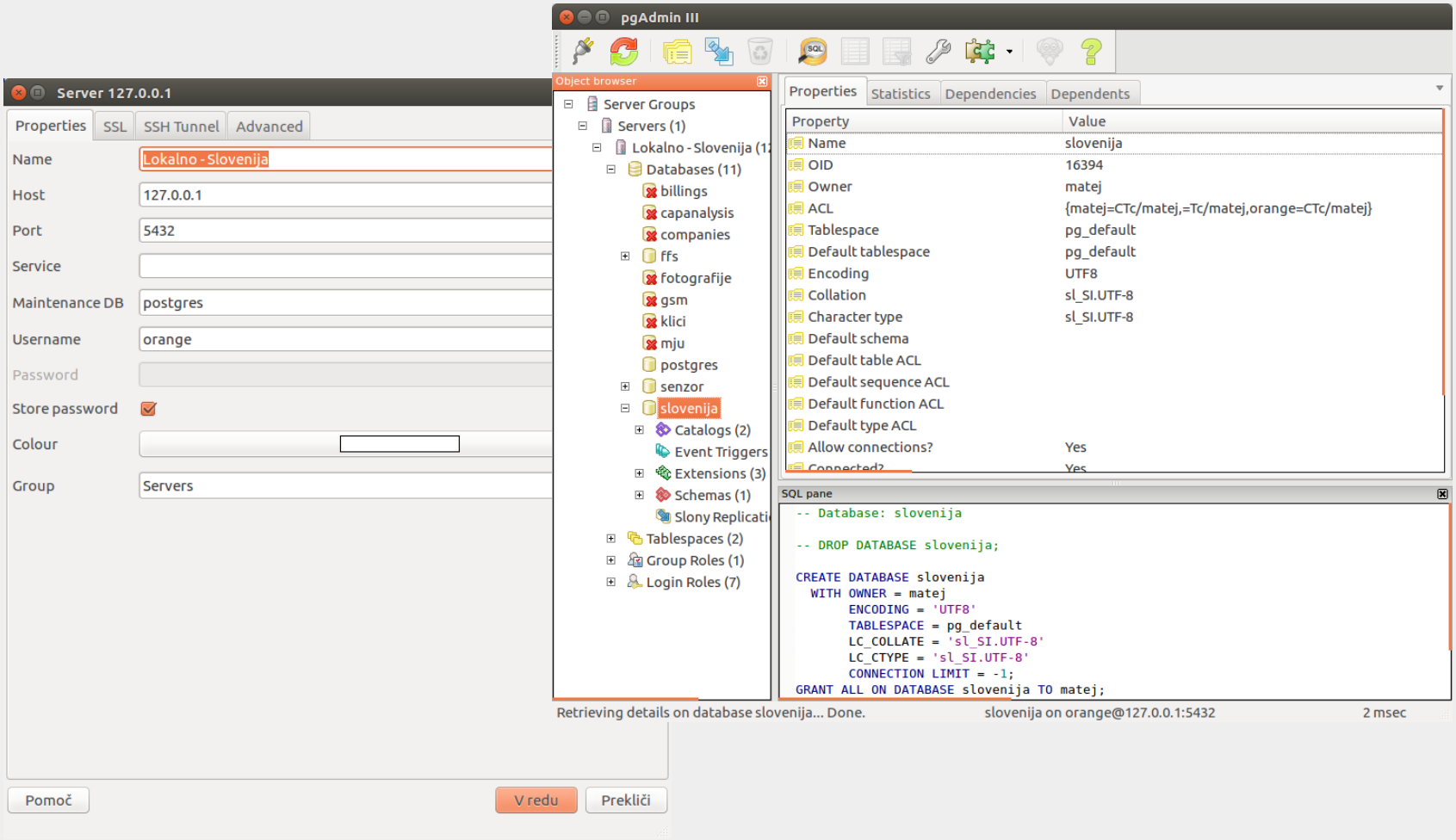
Namestitev je specifična za operacijski sistem.

Po namestitvi je smiselno nekaj časa posvetiti nastavitvi ustreznih pravic dostopa do baz podatkov oz. zavarovanju dostopa do podatkov, zlasti, če bo podatkovni strežnik odprt na internet:

- ustvarjanje novih uporabnikov,
- nastavitev gesel,
- omejitev pravic uporabnikom,
- požarni zid,...

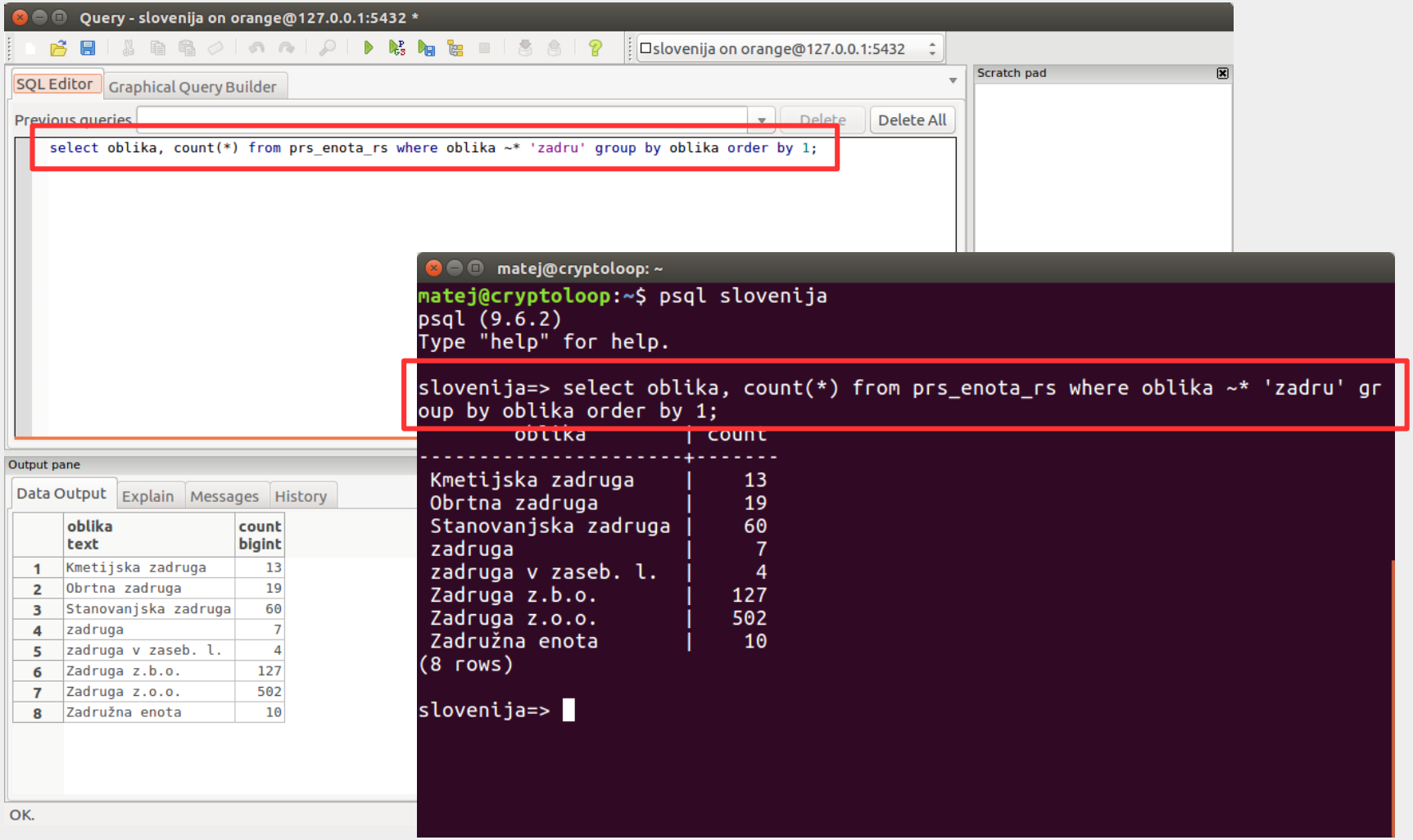
PgAdmin

PgAdmin je grafično orodje za delo s PostgreSQL.
Povezujemo se lahko tudi na oddaljene baze.



PgAdmin vs. ukazna vrstica

Namesto PgAdmin lahko uporabimo tudi ukazno vrstico.



Priprava in uvoz podatkov

Pridobivanje podatkov:

- javni viri,
- zakon o dostopu do informacij javnega značaja.

Razumevanje oblike podatkov:

- CSV, tab-delimited, MDB, XML, JSON...
- kodiranje znakov (izbira tim. kodne tabele znakov, npr. CP-1250, ISO-8859-2, UFT-8).

Priprava in čiščenje podatkov:

- ročno,
- Excel/LibreOffice,
- Google Refine, TextEdit/Gedit, bash...

```
grep -v '^$' podatki.txt >
podatki_brez_praznih_vrstic.txt
```

Ustvarjanje novega uporabnika in nove baze

PgAdmin: *Tools – Query tool* ali:

```
sudo su
su - postgres
psql
```

Ustvarjanje novega uporabnika (v bazi *postgres*):

```
CREATE USER novinar WITH PASSWORD 'transparency';
ALTER USER novinar VALID UNTIL 'infinity';
```

(UTF-8 podpora za vse nove baze.)

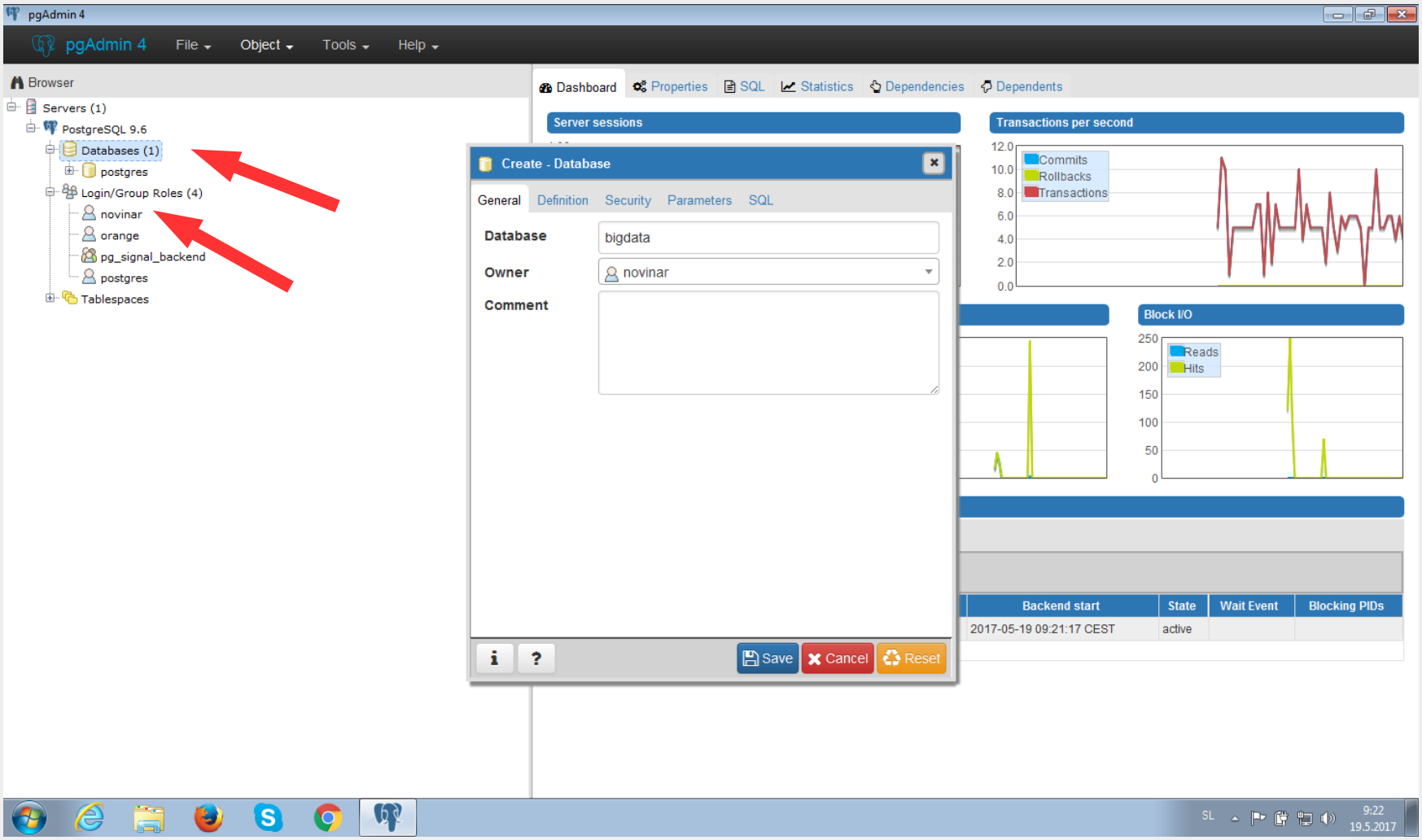
Ustvarjanje nove baze:

```
CREATE DATABASE bigdata WITH OWNER novinar;
```

Dodajanje razširitve za indeksiranje teksta:

```
CREATE EXTENSION pg_trgm;
```

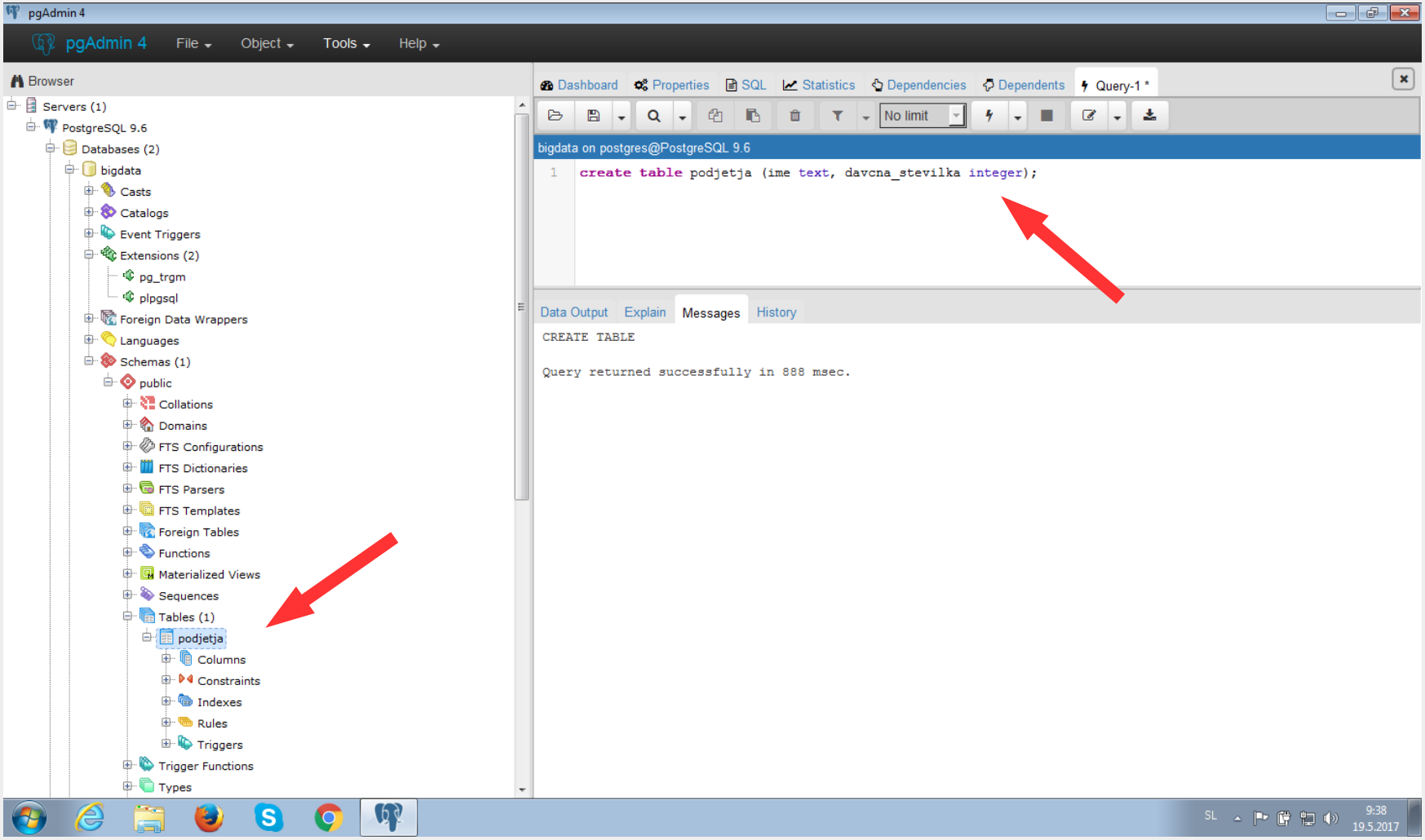
Ustvarjanje novega uporabnika in nove baze



Database - Extensions.

Database - Schemas – Public – Table.

Ustvarjanje novega uporabnika in nove baze



Priprava in uvoz podatkov

Ustvarjanje nove tabele in definiranje njene strukture:

```
create table podjetja (ime text,
davcna_stevilka integer);
```

Zakaj »*text*«?

Uvoz podatkov:

```
copy podjetja (ime, davcna_stevilka) from
'podjetja.txt' with csv header delimiter
E'\t';
```

“Ročno” dodajanje vrstic:

```
insert into podjetja (ime,
maticna_stevilka) values ('podjetje 1',
12345678);
```

Priprava in uvoz podatkov

Ustvarjanje indeksa (za pospešitev operacij nad podatki):

```
CREATE EXTENSION pg_trgm;


CREATE INDEX iime_gist on podjetja USING
gist (ime gist_trgm_ops);
CREATE INDEX idavcna on
podjetja(davcna_stevilka);
```

Dodajanje stolpcev:

```
alter table podjetja add column
davcna_text text;
update podjetja set davcna_text =
davcna_stevilka::text;
```

Primer: DUTB

Ustvarjanje nove tabele in definiranje njene strukture:

```
CREATE TABLE dutb (naziv text,  naslov text,
znesek_text text, banka text, posta text,
postna_stevilka text, maticna text, davcna
integer);
```

Uvoz podatkov:

```
\COPY dutb from 'dutb.txt' with csv header
delimiter E'\t';
```

Dodajanje novega stolpca:

```
alter table dutb add column znesek
numeric(16,2);
update dutb set znesek =
znesek_text::numeric(16,2);
```

Primer: DUTB

Izračun indeksov:

```
CREATE INDEX iznesek on dutb (znesek);
CREATE INDEX idavcna on dutb (davcna);

CREATE INDEX inaziv_gist on dutb USING gist
(naziv gist_trgm_ops);
CREATE INDEX ibanka_gist on dutb USING gist
(banka gist_trgm_ops);
```

Primeri analiz

Pregled podatkov:

-S q

```
select * from dutb;
select naziv from dutb;
select banka from dutb;
select banka, naziv from dutb;

select * from dutb limit 5;

select * from dutb order by znesek;
select naziv, znesek, banka from dutb
order by znesek desc;

select naziv, znesek, banka from dutb
order by 1 desc;
```

Primeri analiz

Pregled podatkov s pogojem:

```
select naziv, znesek, banka from dutb
where (banka = 'NLB');

select naziv, znesek, banka from dutb
where (banka = 'NLB') order by znesek
desc;

select naziv, znesek, banka from dutb
where (banka = 'NLB') and (znesek >
100000);

select naziv, znesek, banka from dutb
where (naziv ~* 'rea');
select naziv, znesek, banka from dutb
where (naziv ~* '^rea');
```

Primeri analiz

Sumarni prikazi:

```
select sum(znesek), banka from dutb group
by banka;
select count(*), banka from dutb group by
banka;
select count(*), banka from dutb where
(znesek > 100000) group by banka;

select min(znesek), banka from dutb group
by banka;

select avg(znesek), banka from dutb group
by banka;
select round(avg(znesek),2), banka from
dutb group by banka;
```

Primeri analiz

Najprej uvozimo Supervizor podatke:

```
create table supervizor (sifra_pu integer,
davcna_stevilka integer, leto integer,
mesec integer, vsota_prejemkov
numeric(16,4), nepovratna_sredstva
numeric(16,4), fiduciarni_posli
numeric(16,4));

\COPY supervizor from 'supervizor_2003-
2014.txt' with csv header delimiter E'\t'
```

Primeri analiz

Izračun indeksov nad Supervizor bazo:

```
CREATE INDEX ivsota_prejemkov_supervizor
on supervizor (vsota_prejemkov);

CREATE INDEX idavcna_stevilka_supervizor
on supervizor (davcna_stevilka);
```

Primeri analiz

Povezovanje več tabel:

```
select * from dutb, supervizor where
(dutb.davcna = supervizor.davcna_stevilka);
```

```
select naziv, sum(znesek) as slabi_krediti, sum(vsota_prejemkov) as
prejemki_drzave from dutb, supervizor where (dutb.davcna =
supervizor.davcna_stevilka) group by naziv order by 2 desc;
```

```
select (select dutb.naziv from dutb where
dutb.davcna = supervizor.davcna_stevilka limit
1) as ime_podjetja, davcna, sum(znesek) as
slabi_krediti, sum(vsota_prejemkov) as
prejemki_drzave from dutb, supervizor where
(dutb.davcna = supervizor.davcna_stevilka)
group by ime_podjetja, davcna order by 2 desc;
```

Primeri analiz

Izpis rezultatov v datoteko:

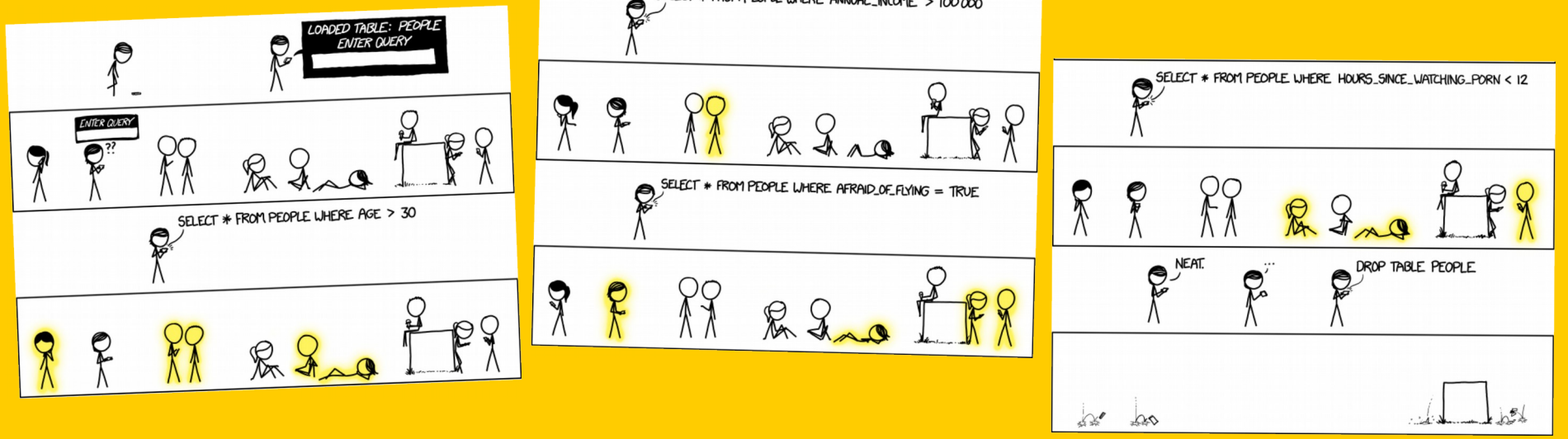
```
\COPY (select count(*), banka from dutb
where (znesek > 100000) group by banka) to
'banke_nad100k.txt' with csv header
delimiter E'\t';
```

Izbris tabele (previdno!):

```
DROP TABLE dutb;
```

Ctrl-d

SELECT *questions* FROM *audience*;



Cartoon: (CC) xkcd.com

Matej Kovačič
matej.kovacic@ijs.si

Personal blog: https://pravokator.si