THE UNIVERSITY OF CHICAGO

MindBytes 2018

Irene Farah
Center for Spatial Data Science

THE CENTER FOR SPATIAL DATA SCIENCE
THE UNIVERSITY OF CHICAGO

Building an automated and replicable PostGIS data warehouse for U.S. business establishments

Objectives

- **Create a PostGIS spatial database on the University of Chicago's Research Computing Center (RCC) server.**
- **Create a replicable automated process for building a spatial database.**
- **Optimize the import process.**
- **Create a replicable automated process for running queries.**

Specifics:

- **Import the data into the spatial database (using \copy command)**
- **Parallelize the import process (using Parallel GNU)**
- **Run replicable and efficient queries using the NETS (National Establishments Time-Series) data with ~ 60 million datapoints.**
- **Show an application of the filtered data.**

Methodology

Build data warehouse

1

- **CREATE spatial_db**
- **CREATE spatial extensions**
  - ★*postgis*
  - ★*postgis_topology*
  - ★*hstore*
  - ★*pgRouting*

2 CREATE database and tables

COPY scp .txt files → **la2.rcc.uchicago.edu**

NETS Data Tables

Table	GB	Observations
HQ Company	17.35	58,861,745
Company	16.62	58,861,745
SIC	11.25	58,861,745
Misc. (spatial)	9.07	58,861,745
Sales	7.32	58,861,745
Address	7.12	58,861,745
Ratings	6.64	58,861,745
Headquarters	6.33	58,861,745
Employment	5.21	58,861,745
NAICS	4.76	58,861,745
FIPS	4.57	58,861,745
Move	2.31	5,280,005
MoveSummary	0.73	4,435,243

split
Company and HQ Company into 16 tables

> 11.25 GB

< 11.25 GB

3 Import: load input

\copy

CREATE SCHEMA
CREATE TABLES
Load data

PARALLEL GNU

Time for importing 45 tables

- time parallel -j8 -a script eval
  time: **12 minutes** 38s
- One by one
  time: **43 minutes** 67s

4 Clean tables

- **UNION** of split tables
- **CREATE INDEXES**
- **Lowercase** columns in schema
- **Convert** data types

5 Create spatial environment

- Define **spatial reference system**
- Add **geometry type (points)** from table that contains **latitude** and **longitude**
- **Create spatial index**

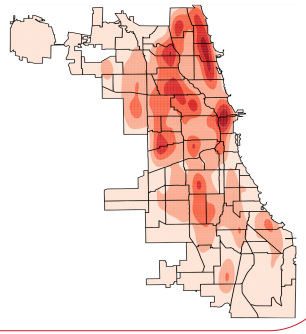
Applications

Query datapoints

- Automated query for Chicago's **retail** from 1990 – 2014 (**join** and filter with **select**)
- Filter for small independent business groceries* in a replicable and automated script.
- Export from server → local file (using **psql2shp)**
- Calculate the **Kernel Density Estimation** in R
  → Greatest concentration of small and independent grocery shops is in the center and north part of Chicago.

** Small independent business groceries are defined as **standalone** establishments with less than **10 employees.***

Figure 1. Kernel Density Estimation Map of Chicago's Small Business Groceries



Conclusions

- **Capacity limit** for loading data into PostgreSQL is ~ 11GB.
- Use **split** if tables exceed limit for optimal performance.
- **Parallelizing** the import improves performance by ~4x, using the 8 CPU cores from the server.
- **Parallel GNU** provides better control and is more efficient than **xargs** when using the server (keeps CPUs active, generating a new process when one finishes).
- Parallel **queries** for the 24 years in the NETS data also makes filtering more efficient and enables more manageable data analysis across software.

Source: Walls & Associates (2014). National Establishment Time-Series (NETS) Database.
Special thanks for contributions and guidance to the Center for Spatial Data Science and the RCC team.

May 2018 - irenef@uchicago.edu