Enrich your data with geocoordinates from OpenStreetMap or ArcGIS

Andreas 'ads' Scherbaum
Andreas Scherbaum

• Works with databases since 1997, with PostgreSQL since 1998
• Founding member of PGEU
• Board of Directors: PGEU
• PostgreSQL Regional Contact for Germany
• Ran my own company around PostgreSQL for 7+ years
• Joined EMC in 2011, working on PostgreSQL and Greenplum projects
Get Social

• Google+: http://goo.gl/e3EtB1
• Twitter: @ascherbaum
Table of Content

• Why (enrich your data)
• How (to enrich your data)
• Measure (the results)
• What’s next
• #icecream
Why (enrich your data)
Geolocation Services

- Identify customer location
  - Country, Region, City, Street
- Fraud Prevention
- Geo Marketing
- Geo Tagging
- Geo Targeting
- Value Added Tax in the European Union
  - Tax is paid based on customer location
- Weather Forecast
- Concurrency Calculation
- ...

Geolocation Services

• I work for a Data Warehouse company ...
  o Many of our customers have millions of addresses in the EDW
Let’s Talk Business

Let’s imagine, you run a small business
  • A pizza shop
  • State-of-the-art website
  • Online orders

Your claim:
  • Always hot pizza, or money back
Berlin
Germany!
Do you know your customers?

• Where is:
  • Uthlande Street 1, **25849 Pellworm**, Germany
  • (address of the local government)
Do you know your customers?

- Where is:
  - Uthlande Street 1, **25849 Pellworm**, Germany
  - (address of the local government)

- It's in the „Northern Friesland“ district

© Wikipedia
Do you know your customers?

- Where is:
  - Uthlande Street 1, **25849 Pellworm**, Germany
  - (address of the local government)

- Only reachable via ferry
  - Good luck with your hot pizza in the evening
Do you know your customers?

- Where is:
  - Junkerstraße 86, **78266 Büsingén am Hochrhein**, Germany
  - (address of the local government)
Do you know your customers?

- Where is:
  - Junkerstraße 86, 78266 Büsingen am Hochrhein, Germany
  - (address of the local government)

- It's entirely surrounded by the Swiss canton of Schaffhausen (enclave)
  - Food delivery from Germany to Switzerland has to pass customs

© Wikipedia
Do you know your customers?

- Where is:
  - Ullasvej 23, 3700 Rønne, Denmark
  - (address of the local government)
Do you know your customers?

- Where is:
  - Ullasvej 23, 3700 Rønne, Denmark
  - (address of the local government)

- It’s on Bornholm
  - Danish island in the Baltic Sea
  - Google says: need a ferry
  - and 4.5 hours

© Wikipedia
Who cares?

• No one cares about corner cases, right?
• Let‘s deal with that later ...
  • Yeah, always a good idea, especially in IT
Let’s start easy

• Zip codes, that works everywhere, every street has one
• Germany uses 5 digits for zip codes
  • Other countries use 4, or 6, or characters
• But note: close number‘s don‘t imply close distance
Geography of Germany

• Wikipedia: „The territory of Germany covers 357,021 km²...”

\[
\frac{x}{1 \text{ zip code}} = \frac{357021 \text{ km}^2}{99999 \text{ zip code}}
\]

\[
x = \frac{357021 \text{ km}^2}{99999 \text{ zip code}} = 3.57 \text{ km}^2
\]
Geography of Germany

• Wikipedia:
  o northernmost point in mainland Germany: Aventoft, Schleswig-Holstein
    $54°54′N \ 8°49′E$
  o southernmost point: Haldenwanger Eck, Oberstdorf, Bavaria
    $47°16′12.39″N \ 10°10′41.95″E$

SELECT ST_Distance_Sphere( ST_GeomFromText( 'POINT(54.900176 8.816667)', 4326 ), ST_GeomFromText( 'POINT(47.270108 10.178319)', 4326 )) / 1000;

850.346457590326
(1 row)

\[
\frac{x}{1 \text{ zip code}} = \frac{850 \text{ km}}{99999 \text{ zip code}} \quad \Rightarrow \quad x = \frac{850 \text{ km}}{99999 \text{ zip code}} = 0.0085 \text{ km}
\]
Let's start easy

- Question: what is the distance between:
  - 08523 Plauen, Germany
  - 95028 Hof (Saale), Germany

\[
x = \frac{86505 \text{ zip code}}{850 \text{ km}} = 735.4 \text{ km}
\]

\[
x = \frac{850 \text{ km} \times 86505 \text{ zip code}}{99999 \text{ zip code}} = 735.4 \text{ km}
\]
Let’s start easy

- Question: what is the distance between:
  - **08523** Plauen, Germany
  - **95028** Hof (Saale), Germany

- Merely 42 km

© Google Maps
Ok ...

- That escalated quickly ...
- Zip codes aren’t easy
- Ferry doesn’t work, and neither does customs

$ psql
psql (9.3.4)
Type "help" for help.
Time ...

- to GeoCode addresses
Geocoding

- Look up geocoordinates
  - Longitude: X
  - Latitude: Y
Reverse Geocoding

• Look up near(est) address for geocoordinates
How (to enrich your data)
Data Sources

- OpenStreetMap
- ArcGIS
- Google Maps
- Bing Maps
- ...

OpenStreetMap

• „OpenStreetMap is built by a community of mappers that contribute and maintain data about roads, trails, cafés, railway stations, and much more, all over the world.‟
  o Quote from http://www.openstreetmap.org/about

• Project „Nominatim“ provides a webservice for Geocoding and Reverse Geocoding

• You can (and should) setup your own server
ArcGIS

- Commercial service, offering maps, web platform, app framework, paid services and much more
- Also provides free Geocode and Reverse Geocode in the web API
PostgreSQL

- Let’s do this from with the database:
  - **pg_osm**: [https://github.com/andreasscherbaum/pg_osm](https://github.com/andreasscherbaum/pg_osm)
  - **pg_arcgis**: [https://github.com/andreasscherbaum/pg_arcgis](https://github.com/andreasscherbaum/pg_arcgis)

```sql
arcgis=# SELECT * from arcgis.find('3495 Deer Creek Road Palo Alto, CA 94304');
NOTICE: resolving: 3495 Deer Creek Road Palo Alto, CA 94304
CONTEXT: PL/Python function "find"

 x       |       y       | wkid
---------------------+---------------------
-122.148659529 | 37.3939849106 | 4326

(1 row)
```
PostgreSQL

arcgis=# select arcgis.find_yx('3495 Deer Creek Road Palo Alto, CA 94304');
find_yx

--------------------------------

{37.3939849106, -122.148659529}

(1 row)
find_yx(lat, lon)
Measure (the results)
First demo set

• This started as a project for Pivotal Greenplum Database (a PostgreSQL fork) - performance
• Pivotal has 17 offices worldwide
• Scrape the addresses from the website and load them into the database
DEMO?
Pivotal addresses

- **ArcGIS**: worked out of the box (just removed „floor“, „suite“ etc) for 15 out of 17 addresses
  - China and South Korea are listed for weak data

```sql
arcgis=# select address, location_tmp from test_arcgis.pivotal_addresses order by id;
```

<table>
<thead>
<tr>
<th>address</th>
<th>location_tmp</th>
</tr>
</thead>
<tbody>
<tr>
<td>3495 Deer Creek Road Palo Alto, CA 94304</td>
<td>-122.148659529 37.3939849106</td>
</tr>
<tr>
<td>38 Xiaoyun Road, Beijing, 100027 China</td>
<td>-8.405141449 51.8861076</td>
</tr>
<tr>
<td>City Gate, Mahon, Cork, Ireland</td>
<td>8.532926108 50.1649573</td>
</tr>
<tr>
<td>Am Kronberger Hang 2A, Schwalbach Frankfurt 65824, Germany</td>
<td>-2.61271281 51.225472175</td>
</tr>
<tr>
<td>Bentima House, 168-172 Old Street, London E11V 9BP, United Kingdom</td>
<td>-3.591640162 40.462199799</td>
</tr>
<tr>
<td>Ribera del Loira 8, Edificio Paris, Campo de las Naciones, Madrid 28042, Spain</td>
<td>9.190325913 45.443185172</td>
</tr>
<tr>
<td>Via Spadolini, 5, Edificio A, Milano 20141, Italy</td>
<td>72.859927111 19.062134031</td>
</tr>
<tr>
<td>2000 Bandra Kurla Complex, Bandra East, Mumbai, Maharashtra, 400051, IND</td>
<td>-73.9948157784 40.740185687</td>
</tr>
<tr>
<td>625 Avenue of the Americas, New York, NY 10011-2020, United States</td>
<td>2.2202362 48.9228068</td>
</tr>
<tr>
<td>80 Quai Voltaire, Bezons Cedex 95876, France</td>
<td>151.193129783 33.823683444</td>
</tr>
<tr>
<td>875 Howard St, San Francisco, CA 94103, United States</td>
<td>139.698956933 35.686583575</td>
</tr>
<tr>
<td>152 Teheran-ro, Seoul, 135-984, Korea</td>
<td>139.698956933 35.686583575</td>
</tr>
<tr>
<td>1 Changi Business Park Central 1, #08-101, Changi City, Singapore 486036</td>
<td>103.962589042 1.33351667</td>
</tr>
<tr>
<td>207 Pacific Highway, St Leonards, Sydney, NSW, 2065, Australia</td>
<td>139.698956933 35.686583575</td>
</tr>
<tr>
<td>2-1-1 Yoyogi Shibuya-ku, Tokyo 151-0053, Japan</td>
<td>-79.376037693 43.64983575</td>
</tr>
<tr>
<td>1 Toronto Street, Suite 1100, Toronto, Ontario, M5C 2V6, Canada</td>
<td>5.098633003 52.03586606</td>
</tr>
</tbody>
</table>

(17 rows)
**Pivotal addresses**

- **OSM: out of the box, only 4 addresses resolved**

```sql
arcgis=# select address,location_tmp from test_osm.pivotal_addresses order by id;

<table>
<thead>
<tr>
<th>address</th>
<th>location_tmp</th>
</tr>
</thead>
<tbody>
<tr>
<td>3495 Deer Creek Road Palo Alto, CA 94304</td>
<td>-122.14782867439 37.394282</td>
</tr>
<tr>
<td>38 Xiaoyun Road, Beijing, 100027 China</td>
<td>116.4616591 39.9602201</td>
</tr>
<tr>
<td>City Gate, Mahon, Cork, Ireland</td>
<td></td>
</tr>
<tr>
<td>Am Kronberger Hang 2A, Schwalbach Frankfurt 65824, Germany</td>
<td></td>
</tr>
<tr>
<td>Bentima House, 168-172 Old Street, London E1 9BP, United Kingdom</td>
<td></td>
</tr>
<tr>
<td>Ribera del Loira 8, Edificio Paris, Campo de las Naciones, Madrid 28042, Spain</td>
<td></td>
</tr>
<tr>
<td>Via Spadolini, 5, Edificio A, Milano 20141, Italy</td>
<td></td>
</tr>
<tr>
<td>2000 Bandra Kurla Complex, Bandra East, Mumbai, Maharashtra, 400051, IND</td>
<td></td>
</tr>
<tr>
<td>625 Avenue of the Americas, New York, NY 10011-2020, United States</td>
<td></td>
</tr>
<tr>
<td>80 Quai Voltaire, Bezons Cedex 95876, France</td>
<td></td>
</tr>
<tr>
<td>875 Howard St, San Francisco, CA 94103, United States</td>
<td>-122.4137689 37.774516</td>
</tr>
<tr>
<td>152 Teheran-ro, Seoul, 135-984, Korea</td>
<td></td>
</tr>
<tr>
<td>1 Changi Business Park Central 1, #08-101, Changi City, Singapore 486036</td>
<td></td>
</tr>
<tr>
<td>207 Pacific Highway, St Leonards, Sydney, NSW, 2065, Australia</td>
<td></td>
</tr>
<tr>
<td>2-1-1 Yoyogi Shibuya-ku, Tokyo 151-0053, Japan</td>
<td>151.1883288 -33.8245562</td>
</tr>
<tr>
<td>1 Toronto Street, Suite 1100, Toronto, Ontario, M5C 2V6, Canada</td>
<td></td>
</tr>
<tr>
<td>Edisonbaan 14b , 3439 MN Nieuwegein, 3430 AB Nieuwegein, Netherlands PO Box 97, Netherlands</td>
<td></td>
</tr>
</tbody>
</table>
```

(17 rows)
**Pivotal addresses**

- **OSM:** with „address tuning“, max 10 addresses resolved
  - But: China and South Korea resolved!

```
arcgis=# select address,location_tmp from test_osm.pivot_addresses order by id;

<table>
<thead>
<tr>
<th>address</th>
<th>location_tmp</th>
</tr>
</thead>
<tbody>
<tr>
<td>3495 Deer Creek Road Palo Alto, CA 94304</td>
<td>-122.14782867439 37.394282</td>
</tr>
<tr>
<td>38 Xiaoyun Road, Beijing, 100027 China</td>
<td>116.4616591 39.9602201</td>
</tr>
<tr>
<td>City Gate, Mahon, Cork, Ireland</td>
<td></td>
</tr>
<tr>
<td>Am Kronberger Hang 2A, Schwalbach 65824, Germany</td>
<td>8.5330088 50.1643734</td>
</tr>
<tr>
<td>168-172 Old Street, London EC1V 9BP, United Kingdom</td>
<td>-0.0905437 51.5251618</td>
</tr>
<tr>
<td>Ribera del Loira 8, Edificio Paris, Campo de las Naciones, Madrid 28042, Spain</td>
<td></td>
</tr>
<tr>
<td>Via Spadolini, 5, Edificio A, Milano 20141, Italy</td>
<td></td>
</tr>
<tr>
<td>2000 Bandra Kurla Complex, Bandra East, Mumbai, Maharashtra, 400051, IND</td>
<td></td>
</tr>
<tr>
<td>625 Avenue of the Americas, New York, NY 10011, United States</td>
<td>-73.9949875 40.7402113</td>
</tr>
<tr>
<td>80 Quai Voltaire, Bezons Cedex 95876, France</td>
<td></td>
</tr>
<tr>
<td>875 Howard St, San Francisco, CA 94103, United States</td>
<td>-122.4137689 37.774516</td>
</tr>
<tr>
<td>152 Teheran-ro, Seoul, 135, South Korea</td>
<td>127.063216 37.5087407</td>
</tr>
<tr>
<td>1 Changi Business Park Central 1, #08-101, Changi City, Singapore 486036</td>
<td></td>
</tr>
<tr>
<td>207 Pacific Highway, St Leonards, Sydney, NSW, 2065, Australia</td>
<td>151.1883288 -33.8245562</td>
</tr>
<tr>
<td>2-1-1 Yoyogi Shibuya-ku, Tokyo 151-0053, Japan</td>
<td></td>
</tr>
<tr>
<td>1 Toronto Street, Toronto, Canada</td>
<td>-79.3759569 43.6500879</td>
</tr>
<tr>
<td>Edisonbaan 14b, Nieuwegein, Netherlands</td>
<td>5.0963629 52.034968</td>
</tr>
</tbody>
</table>
(17 rows)
```
Second demo set

- Pivotal is a spin-off from EMC and VMware
  - Let's look after EMC addresses
- Yielded 196 addresses worldwide
  - Excluding USA (another test set)
  - Excluding some addresses which not even Google Maps could resolve
EMC has funny offices

© Google Maps
Second demo set

- EMC addresses worldwide (non USA)
  - 196 addresses in data set
  - ArcGIS: 165 resolved
  - OSM: 71 resolved
    - OSM found 3 places which ArcGIS couldn’t find
Second demo set

- EMC addresses worldwide (non USA)
  - 196 addresses in data set
  - ArcGIS: 165 resolved (84%)
  - OSM: 71 resolved (36%)
    - OSM found 3 places which ArcGIS couldn't find
  - 68 locations resolved from both services (34%)
    - min distance: 6.7 meters
    - max distance: 6177.34 meters (6.1 km)
    - avg distance: 539.44 meters
    - stddev: 1063.97 (wide spread)
3rd & 4th data set

• 3rd data set: EMC offices USA
  o 216 addresses in total
  o ArcGIS: 216 resolved (100%)
  o OSM: 133 resolved (61%)

• 4th data set: Target shops USA
  o 1779 addresses in total
  o ArcGIS: 1779 resolved (100%)
  o OSM: 1205 resolved (67%)
Conclusion

• ArcGIS:
  o 100% resolve rate in USA
  o 80% - 90% resolve rate outside the USA

• OpenStreetMap:
  o 60% - 70% resolve rate in USA
  o Drops below 40% outside the USA
  o Good data available where ArcGIS is weak
  o In general more data available (if address found)
  o Returned data is not formalized (returns whatever user entered for the data point) – makes it hard to parse
Performance

- Lookups happen per row
  - In PostgreSQL, only one connection is serving the request -> everything is serialized
  - In Greenplum, every segment can do the lookup in parallel
Performance Greenplum

Master Server: Talks SQL to the client

1 – n Segment Server: „hidden“ MPP database

Row processing happens here, in parallel
Performance Greenplum

GPDB Standard Module

27.5TB Data (Usable Physical Capacity per Module)

110TB Useable User Data Capacity per module, 4x compression

Four 2U servers per module, each server contains:
- 2 Sandybridge sockets/16 cores
- 64GB Mem
- 24x 900GB SAS 10K Drives

8 Segmente je Server
- 32 Segmente je Module
- 128 Segmente je Rack
- = 128 parallele Requests
What’s next
Python

- Add proper http headers
- Better error handling
- Configuration (code already there, not used)
PostgreSQL

- Packages
- More test data
- Test „Address Standardizer“
ArcGIS

• Integrate (some) paid services
  o Route calculation
  o Route plotting on map
OpenStreetMap

- Integrate routing service
  - Announced a few weeks ago
What about the #icecream?
#icecream

- The Nordic staff forced me to talk about ice cream

19:43:46 [freenode] [*censored*] oh
19:43:49 [freenode] [*censored*] AND ADD ICECREAM!
#icecream

- 437 pictures with ice cream
  - 137 pictures with GPS coordinates (Exif data)
- 2534 Foursquare Check-ins
  - 857 different locations, with GPS geocoordinates
  - Unknown number of „ice cream“ shops (so far)

- … there are some interesting stats
- But that will be another talk
Thank you