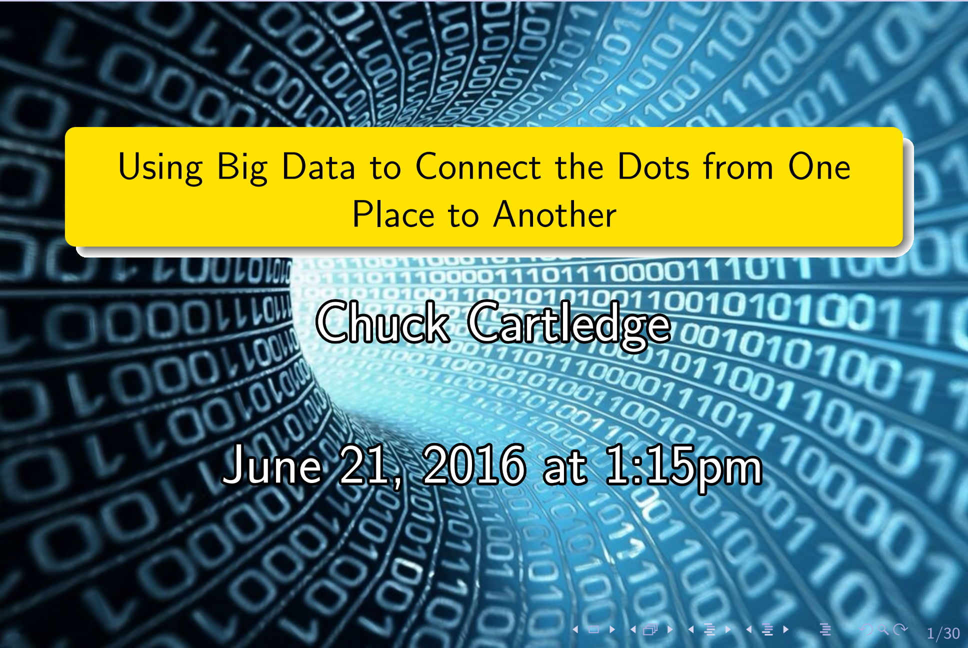


Using Big Data to Connect the Dots from One Place to Another

Chuck Cartledge

June 21, 2016 at 1:15pm

Table of contents I

Königsberg and Euler [1]

- Popular Sunday puzzle was to cross all bridges exactly once
- Mayor asks Leonard Euler to solve the puzzle
- Euler declines, thought the problem was trivial
- Euler changes mind.
- Euler publishes paper in 1736 detailing impossible solution, formulates general solution

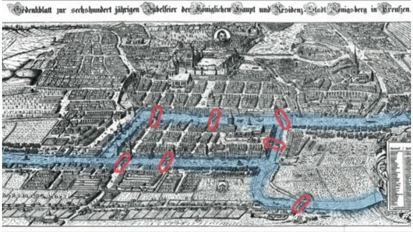
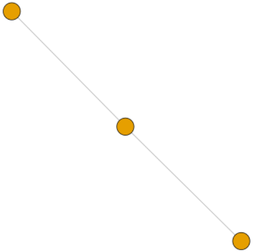

Gedenkblatt zur sechshundert jährigen Jubelfeier der Königlichen Haupt und Residenz-Stadt Königsberg in Preußen.

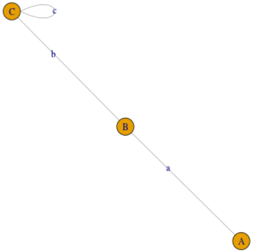
Image from [1].

Birth place of graph theory.

Types of graphs (1 of 2)

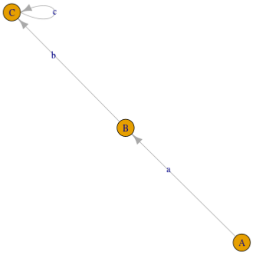


(a) Simple

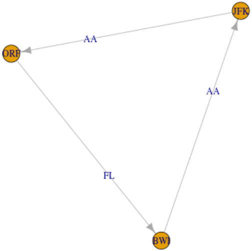


(b) Labeled

Types of graphs (2 of 2)

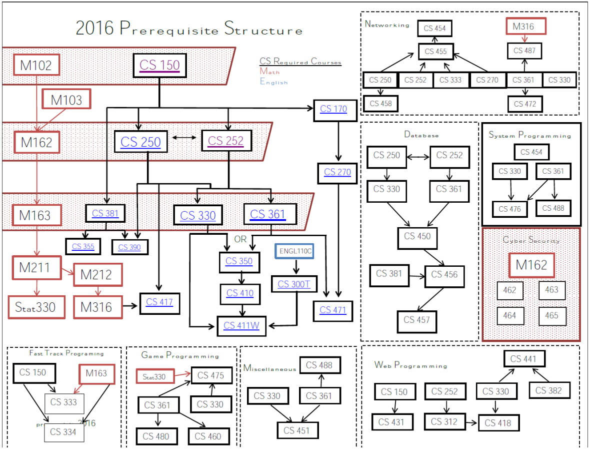


(c) Directed



(d) Useful

ODU Computer Science undergraduate prerequisites



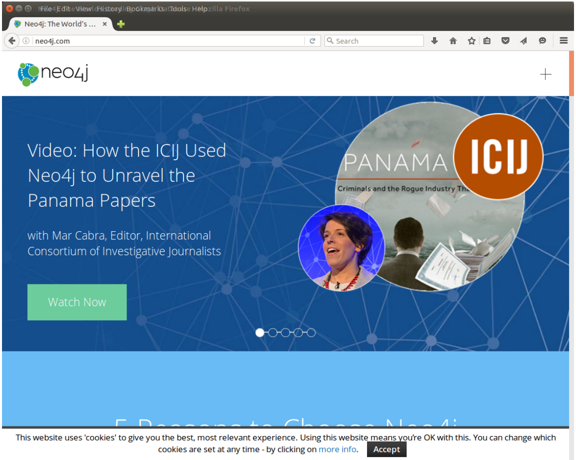
Terms and definitions

- nodes ≡ vertices
- arcs ≡ edges
- arcs ≡ relationships

Terms used interchangeably.

Neo4j uses the terms nodes and relationships.

Where to get it.


File Edit View History Bookmarks Tools Help
Neo4j: The World's ...
neo4j.com
Search
neo4j
Watch Now
PANAMA
Criminals and the Rogue Industry Th
ICIJ
Accept

Video: How the ICIJ Used Neo4j to Unravel the Panama Papers

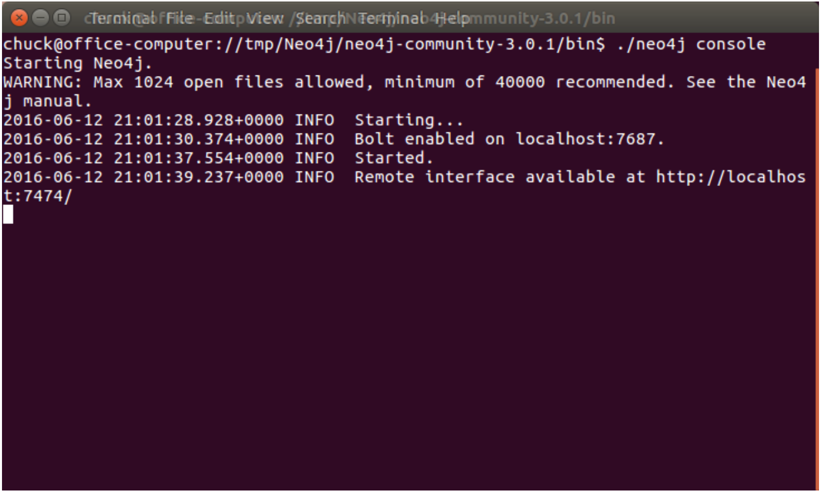
with Mar Cabra, Editor, International Consortium of Investigative Journalists

5 Reasons to Choose Neo4j

This website uses 'cookies' to give you the best, most relevant experience. Using this website means you're OK with this. You can change which cookies are set at any time - by clicking on more info.

Figure: Available with different licenses from `http://neo4j.com/`

How to start it.

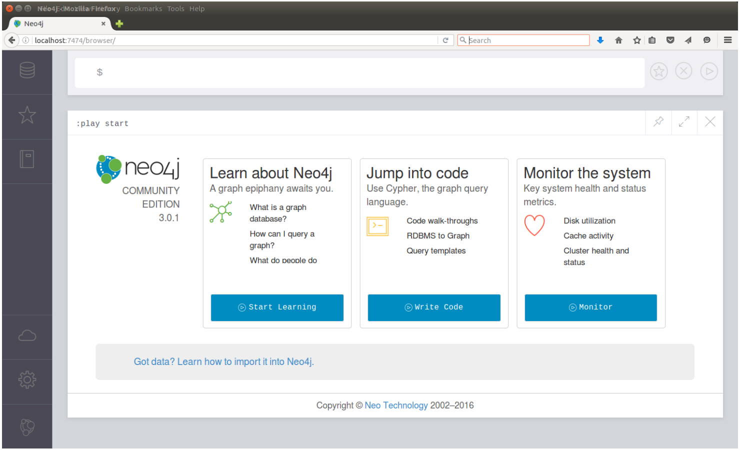


Terminal File Edit View Search Terminal Help
chuck@office-computer: /tmp/Neo4j/neo4j-community-3.0.1/bin

```
chuck@office-computer://tmp/Neo4j/neo4j-community-3.0.1/bin$ ./neo4j console
Starting Neo4j.
WARNING: Max 1024 open files allowed, minimum of 40000 recommended. See the Neo4
j manual.
2016-06-12 21:01:28.928+0000 INFO  Starting...
2016-06-12 21:01:30.374+0000 INFO  Bolt enabled on localhost:7687.
2016-06-12 21:01:37.554+0000 INFO  Started.
2016-06-12 21:01:39.237+0000 INFO  Remote interface available at http://localhos
t:7474/
```

Figure: Starting neo4j manually.

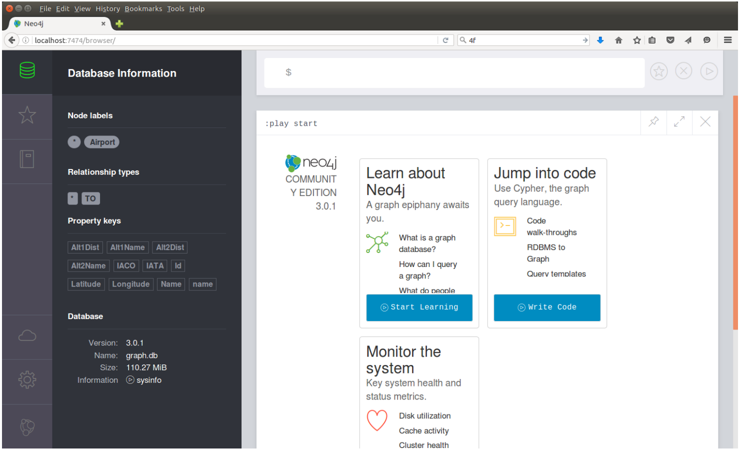
How to interact with it (1 of 4).



Neo4j - Mozilla Firefox Bookmarks Tools Help
Neo4j
localhost:7474/browser/
Search
$
:play start
neo4j
COMMUNITY EDITION
3.0.1
Learn about Neo4j
A graph epiphany awaits you.
What is a graph database?
How can I query a graph?
What do people do
Start Learning
Jump into code
Use Cypher, the graph query language.
Code walk-throughs
RDBMS to Graph
Query templates
Write Code
Monitor the system
Key system health and status metrics.
Disk utilization
Cache activity
Cluster health and status
Monitor
Got data? Learn how to import it into Neo4j.
Copyright © Neo Technology 2002–2016

Figure: Via a browser. Note the URL.

How to interact with it (2 of 4).



Database Information

Node labels

* Airport

Relationship types

* TO

Property keys

Alt1Dist Alt1Name Alt2Dist Alt2Name IACO IATA Id Latitude Longitude Name name

Database

Version:	3.0.1
Name:	graph.db
Size:	110.27 MiB
Information	sysinfo

```
$
```

```
:play start
```

neo4j COMMUNITY EDITION 3.0.1

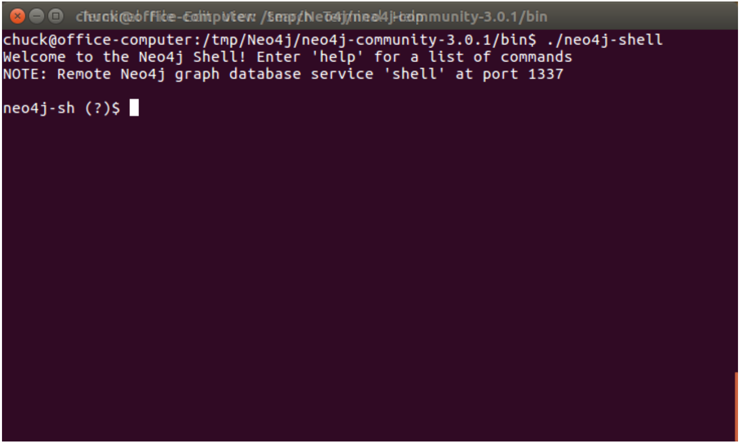
Learn about Neo4j
A graph epiphany awaits you.
What is a graph database?
How can I query a graph?
What do people

Jump into code
Use Cypher, the graph query language.
Code walk-throughs
RDBMS to Graph
Query templates

Monitor the system
Key system health and status metrics.
Disk utilization
Cache activity
Cluster health

Figure: Information about the database.

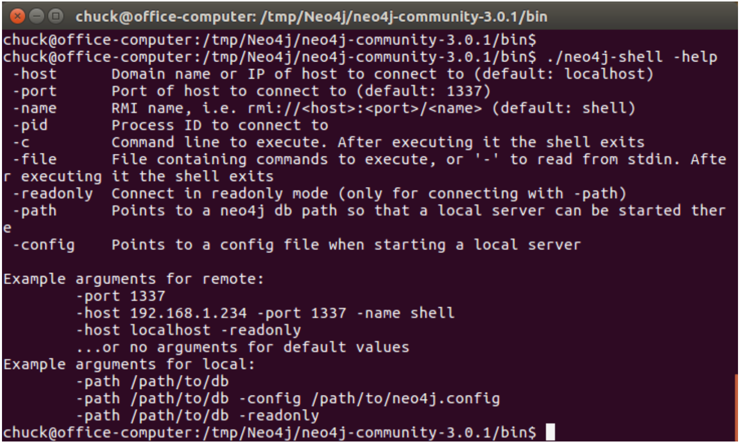
How to interact with it (3 of 4).



```
chuck@office-computer:/tmp/Neo4j/neo4j-community-3.0.1/bin$ ./neo4j-shell
Welcome to the Neo4j Shell! Enter 'help' for a list of commands
NOTE: Remote Neo4j graph database service 'shell' at port 1337

neo4j-sh (?)$ 
```

Figure: A command line interface.

How to interact with it (4 of 4).



```
chuck@office-computer:/tmp/Neo4j/neo4j-community-3.0.1/bin$
chuck@office-computer:/tmp/Neo4j/neo4j-community-3.0.1/bin$ ./neo4j-shell -help
 -host      Domain name or IP of host to connect to (default: localhost)
 -port      Port of host to connect to (default: 1337)
 -name      RMI name, i.e. rmi://<host>:<port>/<name> (default: shell)
 -pid       Process ID to connect to
 -c         Command line to execute. After executing it the shell exits
 -file      File containing commands to execute, or '-' to read from stdin. Afte
r executing it the shell exits
 -readonly  Connect in readonly mode (only for connecting with -path)
 -path      Points to a neo4j db path so that a local server can be started ther
e
 -config    Points to a config file when starting a local server

Example arguments for remote:
        -port 1337
        -host 192.168.1.234 -port 1337 -name shell
        -host localhost -readonly
        ...or no arguments for default values
Example arguments for local:
        -path /path/to/db
        -path /path/to/db -config /path/to/neo4j.config
        -path /path/to/db -readonly
chuck@office-computer:/tmp/Neo4j/neo4j-community-3.0.1/bin$
```

Figure: Help on the CLI.

Neo4j is idempotent

Neo4j nodes can be thought of as a set where each set member is unique/citemathDictionary.

- The time to insert an unconstrained node can be: $O(n^2)$.
- The time to insert a constrained node can be: $O(n)$.
- The time to execute a query can be: $O(x)$
- The time to re-execute a previous query can be: $O(c)$ (for a small c)

Take away: constrain your nodes.

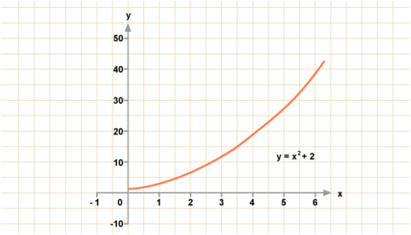


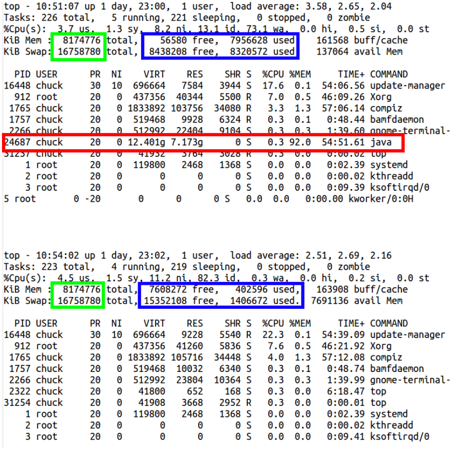
Image from [2].

IMDb (1 of 2)

Table: Time spent loading the IMDb via the neo4j-shell. The time to load the entire IMDb database was too long to be practical.

Size	% IMDb	R-seconds	Neo4j seconds	Nodes loaded
8,000	-	14.070	2.6	2,300
80,000	-	57.462	8.820	26,041
800,000	4	609.890	49.250	198,998
8,000,000	45	1,245.048	556.452	649,309
80,000,000	100	43,600,386.000	1,840,666	1,466,720

IMDB (2 of 2)

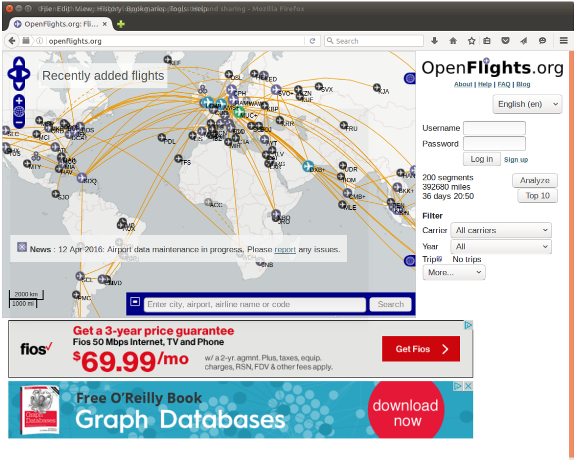


```
top - 10:51:07 up 1 day, 23:00,  1 user,  load average: 3.58, 2.65, 2.04
Tasks: 226 total,   5 running, 221 sleeping,   0 stopped,   0 zombie
%Cpu(s):  3.7 us,  1.3 sy,  8.2 ni, 13.1 id, 73.1 wa,  0.0 hi,  0.5 si,  0.0 st
KiB Mem :  8174776 total,    56580 free,  7956628 used    161568 buff/cache
KiB Swap: 16758780 total,  8438208 free,  8320572 used    137064 avail Mem

  PID USER      PR  NI    VIRT    RES    SHR S  %CPU %MEM     TIME+ COMMAND
16448 chuck     30  10  696664   7584   3944 S  17.6  0.1  54:06.56 update-manager
  912 root      20   0  437356  40344   5500 R   7.0  0.5  46:09.26 Xorg
 1765 chuck     20   0 1833892 103756  34080 R   3.3  1.3  57:06.14 compiz
 1757 chuck     20   0  519468   9928   6324 R   0.3  0.1   0:48.44 bamfdaemon
 2266 chuck     20   0  512992  22404   9104 S   0.3  0.3   1:39.60 gnome-terminal-
24687 chuck     20   0 12.401g 7.173g      0 S   0.3 92.0  54:51.61 java
31237 chuck     20   0   41932   3764   3028 R   0.3  0.0   0:00.02 top
    1 root      20   0  119800   2468   1368 S   0.0  0.0   0:02.39 systemd
    2 root      20   0       0      0      0 S   0.0  0.0   0:00.02 kthreadd
    3 root      20   0       0      0      0 S   0.0  0.0   0:09.39 ksoftirqd/0
5 root       0 -20       0      0      0 S   0.0  0.0   0:00.00 kworker/0:0H
```

```
top - 10:54:02 up 1 day, 23:02,  1 user,  load average: 2.51, 2.69, 2.16
Tasks: 223 total,   4 running, 219 sleeping,   0 stopped,   0 zombie
%Cpu(s):  4.5 us,  1.5 sy, 11.2 ni, 82.3 id,  0.3 wa,  0.0 hi,  0.2 si,  0.0 st
KiB Mem :  8174776 total,  7608272 free,   402596 used,   163908 buff/cache
KiB Swap: 16758780 total, 15352108 free,  1406672 used.  7691136 avail Mem

  PID USER      PR  NI    VIRT    RES    SHR S  %CPU %MEM     TIME+ COMMAND
16448 chuck     30  10  696664   9228   5540 R  22.3  0.1  54:39.09 update-manager
  912 root      20   0  437356  41260   5836 S   7.6  0.5  46:21.92 Xorg
 1765 chuck     20   0 1833892 105716  34448 S   4.0  1.3  57:12.08 compiz
 1757 chuck     20   0  519468  10032   6340 S   0.3  0.1   0:48.74 bamfdaemon
 2266 chuck     20   0  512992  23804  10364 S   0.3  0.3   1:39.99 gnome-terminal-
 2322 chuck     20   0   41800    652    168 S   0.3  0.0   6:18.47 top
31254 chuck     20   0   41908   3668   2952 R   0.3  0.0   0:00.01 top
    1 root      20   0  119800   2468   1368 S   0.0  0.0   0:02.39 systemd
    2 root      20   0       0      0      0 S   0.0  0.0   0:00.02 kthreadd
    3 root      20   0       0      0      0 S   0.0  0.0   0:09.41 ksoftirqd/0
```

Figure: JVM needs swap space.

OpenFlight.org


File Edit View History Bookmarks Tools Help
OpenFlights.org: Fli...
openflights.org
Search
Recently added flights
OpenFlights.org
About | Help | FAQ | Blog
English (en)
Log in
Sign up
Analyze
Top 10
Filter
More...
Enter city, airport, airline name or code
Search
2000 km
1000 mi
Get a 3-year price guarantee
Fios 50 Mbps Internet, TV and Phone
$69.99/mo
w/ a 2-yr. agmnt. Plus, taxes, equip. charges, RSN, FDV & other fees apply.
Get Fios
Free O'Reilly Book
Graph Databases
download now

News : 12 Apr 2016: Airport data maintenance in progress. Please report any issues.

Username

Password

200 segments
392680 miles
36 days 20:50

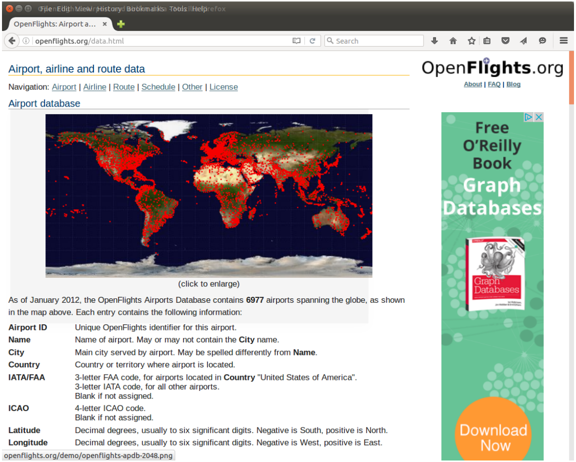
Carrier All carriers

Year All

Trip No trips

Figure: OpenFlight.org homepage. `http://openflights.org/`

OpenFlight.org database.



File Edit View History Bookmarks Tools Help
OpenFlights: Airport a... ×
openflights.org/data.html
Search
(click to enlarge)
openflights.org/demo/openflights-apdb-2048.png

OpenFlights.org

About | FAQ | Blog

Airport, airline and route data

Navigation: Airport | Airline | Route | Schedule | Other | License

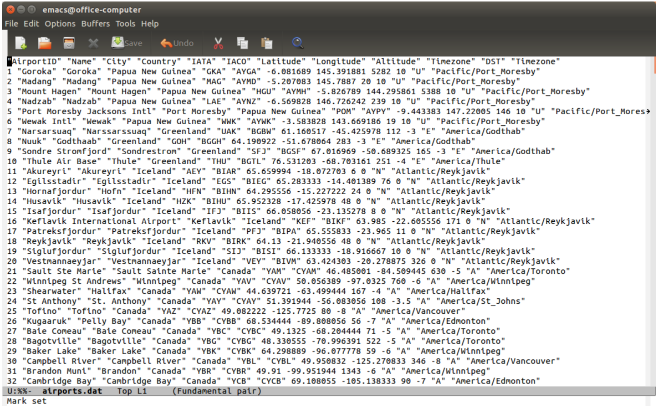
Airport database

As of January 2012, the OpenFlights Airports Database contains **6977** airports spanning the globe, as shown in the map above. Each entry contains the following information:

Airport ID	Unique OpenFlights identifier for this airport.
Name	Name of airport. May or may not contain the **City** name.
City	Main city served by airport. May be spelled differently from **Name**.
Country	Country or territory where airport is located.
IATA/FAA	3-letter FAA code, for airports located in **Country** "United States of America". 3-letter IATA code, for all other airports. Blank if not assigned.
ICAO	4-letter ICAO code. Blank if not assigned.
Latitude	Decimal degrees, usually to six significant digits. Negative is South, positive is North.
Longitude	Decimal degrees, usually to six significant digits. Negative is West, positive is East.

Figure: Interested in airport locations, and service between airports.
http://openflights.org/data.html

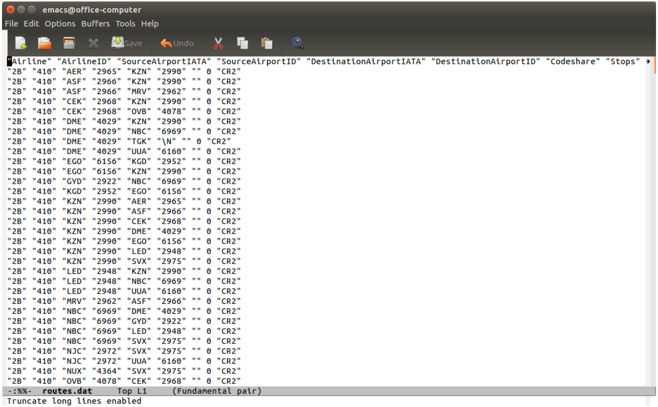
Sample airport data.



```
"AirportID" "Name" "City" "Country" "IATA" "IACO" "Latitude" "Longitude" "Altitude" "Timezone" "DST" "Timezone"
1 "Goroka" "Goroka" "Papua New Guinea" "GKA" "AYGA" -6.081689 145.391881 5282 10 "U" "Pacific/Port_Moresby"
2 "Madang" "Madang" "Papua New Guinea" "MAG" "AYMD" -5.207083 145.7887 20 10 "U" "Pacific/Port_Moresby"
3 "Mount Hagen" "Mount Hagen" "Papua New Guinea" "HGU" "AYMH" -5.826789 144.295861 5388 10 "U" "Pacific/Port_Moresby"
4 "Nadzab" "Nadzab" "Papua New Guinea" "LAE" "AYNZ" -6.569828 146.726242 239 10 "U" "Pacific/Port_Moresby"
5 "Port Moresby Jacksons Intl" "Port Moresby" "Papua New Guinea" "POM" "AYPY" -9.443383 147.22005 146 10 "U" "Pacific/Port_Mores▸
6 "Wewak Intl" "Wewak" "Papua New Guinea" "WWK" "AYWK" -3.583828 143.669186 19 10 "U" "Pacific/Port_Moresby"
7 "Narsarsuaq" "Narssarssuaq" "Greenland" "UAK" "BGBW" 61.160517 -45.425978 112 -3 "E" "America/Godthab"
8 "Nuuk" "Godthaab" "Greenland" "GOH" "BGGH" 64.190922 -51.678064 283 -3 "E" "America/Godthab"
9 "Sondre Stromfjord" "Sondrestrom" "Greenland" "SFJ" "BGSF" 67.016969 -50.689325 165 -3 "E" "America/Godthab"
10 "Thule Air Base" "Thule" "Greenland" "THU" "BGTL" 76.531203 -68.703161 251 -4 "E" "America/Thule"
11 "Akureyri" "Akureyri" "Iceland" "AEY" "BIAR" 65.659994 -18.072703 6 0 "N" "Atlantic/Reykjavik"
12 "Egilsstadir" "Egilsstadir" "Iceland" "EGS" "BIEG" 65.283333 -14.401389 76 0 "N" "Atlantic/Reykjavik"
13 "Hornafjordur" "Hofn" "Iceland" "HFN" "BIHN" 64.295556 -15.227222 24 0 "N" "Atlantic/Reykjavik"
14 "Husavik" "Husavik" "Iceland" "HZK" "BIHU" 65.952328 -17.425978 48 0 "N" "Atlantic/Reykjavik"
15 "Isafjordur" "Isafjordur" "Iceland" "IFJ" "BIIS" 66.058056 -23.135278 8 0 "N" "Atlantic/Reykjavik"
16 "Keflavik International Airport" "Keflavik" "Iceland" "KEF" "BIKF" 63.985 -22.605556 171 0 "N" "Atlantic/Reykjavik"
17 "Patreksfjordur" "Patreksfjordur" "Iceland" "PFJ" "BIPA" 65.555833 -23.965 11 0 "N" "Atlantic/Reykjavik"
18 "Reykjavik" "Reykjavik" "Iceland" "RKV" "BIRK" 64.13 -21.940556 48 0 "N" "Atlantic/Reykjavik"
19 "Siglufjordur" "Siglufjordur" "Iceland" "SIJ" "BISI" 66.133333 -18.916667 10 0 "N" "Atlantic/Reykjavik"
20 "Vestmannaeyjar" "Vestmannaeyjar" "Iceland" "VEY" "BIVM" 63.424303 -20.278875 326 0 "N" "Atlantic/Reykjavik"
21 "Sault Ste Marie" "Sault Sainte Marie" "Canada" "YAM" "CYAM" 46.485001 -84.509445 630 -5 "A" "America/Toronto"
22 "Winnipeg St Andrews" "Winnipeg" "Canada" "YAV" "CYAV" 50.056389 -97.0325 760 -6 "A" "America/Winnipeg"
23 "Shearwater" "Halifax" "Canada" "YAW" "CYAW" 44.639721 -63.499444 167 -4 "A" "America/Halifax"
24 "St Anthony" "St. Anthony" "Canada" "YAY" "CYAY" 51.391944 -56.083056 108 -3.5 "A" "America/St_Johns"
25 "Tofino" "Tofino" "Canada" "YAZ" "CYAZ" 49.082222 -125.7725 80 -8 "A" "America/Vancouver"
26 "Kugaaruk" "Pelly Bay" "Canada" "YBB" "CYBB" 68.534444 -89.808056 56 -7 "A" "America/Edmonton"
27 "Baie Comeau" "Baie Comeau" "Canada" "YBC" "CYBC" 49.1325 -68.204444 71 -5 "A" "America/Toronto"
28 "Bagotville" "Bagotville" "Canada" "YBG" "CYBG" 48.330555 -70.996391 522 -5 "A" "America/Toronto"
29 "Baker Lake" "Baker Lake" "Canada" "YBK" "CYBK" 64.298889 -96.077778 59 -6 "A" "America/Winnipeg"
30 "Campbell River" "Campbell River" "Canada" "YBL" "CYBL" 49.950832 -125.270833 346 -8 "A" "America/Vancouver"
31 "Brandon Muni" "Brandon" "Canada" "YBR" "CYBR" 49.91 -99.951944 1343 -6 "A" "America/Winnipeg"
32 "Cambridge Bay" "Cambridge Bay" "Canada" "YCB" "CYCB" 69.108055 -105.138333 90 -7 "A" "America/Edmonton"
U:%%-  airports.dat   Top L1     (Fundamental pair)
Mark set
```

Figure: Sample data. There are 8,109 airports identified.

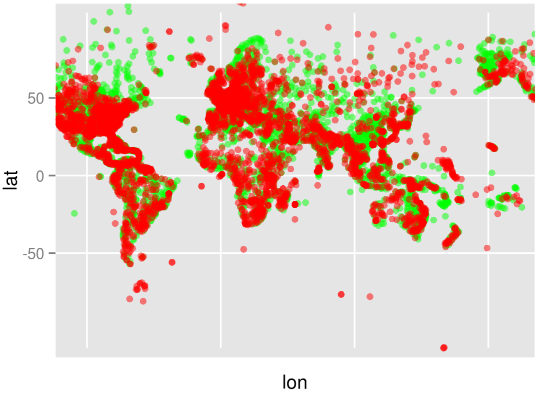
Sample route data.



```
"Airline" "AirlineID" "SourceAirportIATA" "SourceAirportID" "DestinationAirportIATA" "DestinationAirportID" "Codeshare" "Stops"
"2B" "410" "AER" "2965" "KZN" "2990" "" 0 "CR2"
"2B" "410" "ASF" "2966" "KZN" "2990" "" 0 "CR2"
"2B" "410" "ASF" "2966" "MRV" "2962" "" 0 "CR2"
"2B" "410" "CEK" "2968" "KZN" "2990" "" 0 "CR2"
"2B" "410" "CEK" "2968" "OVB" "4078" "" 0 "CR2"
"2B" "410" "DME" "4029" "KZN" "2990" "" 0 "CR2"
"2B" "410" "DME" "4029" "NBC" "6969" "" 0 "CR2"
"2B" "410" "DME" "4029" "TGK" "\N" "" 0 "CR2"
"2B" "410" "DME" "4029" "UUA" "6160" "" 0 "CR2"
"2B" "410" "EGO" "6156" "KGD" "2952" "" 0 "CR2"
"2B" "410" "EGO" "6156" "KZN" "2990" "" 0 "CR2"
"2B" "410" "GYD" "2922" "NBC" "6969" "" 0 "CR2"
"2B" "410" "KGD" "2952" "EGO" "6156" "" 0 "CR2"
"2B" "410" "KZN" "2990" "AER" "2965" "" 0 "CR2"
"2B" "410" "KZN" "2990" "ASF" "2966" "" 0 "CR2"
"2B" "410" "KZN" "2990" "CEK" "2968" "" 0 "CR2"
"2B" "410" "KZN" "2990" "DME" "4029" "" 0 "CR2"
"2B" "410" "KZN" "2990" "EGO" "6156" "" 0 "CR2"
"2B" "410" "KZN" "2990" "LED" "2948" "" 0 "CR2"
"2B" "410" "KZN" "2990" "SVX" "2975" "" 0 "CR2"
"2B" "410" "LED" "2948" "KZN" "2990" "" 0 "CR2"
"2B" "410" "LED" "2948" "NBC" "6969" "" 0 "CR2"
"2B" "410" "LED" "2948" "UUA" "6160" "" 0 "CR2"
"2B" "410" "MRV" "2962" "ASF" "2966" "" 0 "CR2"
"2B" "410" "NBC" "6969" "DME" "4029" "" 0 "CR2"
"2B" "410" "NBC" "6969" "GYD" "2922" "" 0 "CR2"
"2B" "410" "NBC" "6969" "LED" "2948" "" 0 "CR2"
"2B" "410" "NBC" "6969" "SVX" "2975" "" 0 "CR2"
"2B" "410" "NJC" "2972" "SVX" "2975" "" 0 "CR2"
"2B" "410" "NJC" "2972" "UUA" "6160" "" 0 "CR2"
"2B" "410" "NUX" "4364" "SVX" "2975" "" 0 "CR2"
"2B" "410" "OVB" "4078" "CEK" "2968" "" 0 "CR2"
-:%%-  routes.dat     Top L1     (Fundamental pair)
Truncate long lines enabled
```

Figure: Sample route data. There are 67,665 route records. Not all airports have service.

Airports around the world.



Choosing a source and terminating airport.

Available airport data:

- AirportID: Unique OpenFlights identifier for this airport.
- Name: Name of airport. May or may not contain the City name.
- City: Main city served by airport. May be spelled differently from Name.
- Country: Country or territory where airport is located.
- IATA: 3-letter FAA code, for airports located in Country "United States of America". 3-letter IATA code, for all other airports. Blank if not assigned.
- ICAO: 4-letter ICAO code. Blank if not assigned.
- Latitude: Decimal degrees
- Longitude: Decimal degrees
- Altitude: In feet.
- Timezone: Hours offset from UTC
- DST: Daylight savings time
- Timezone: Timezone name

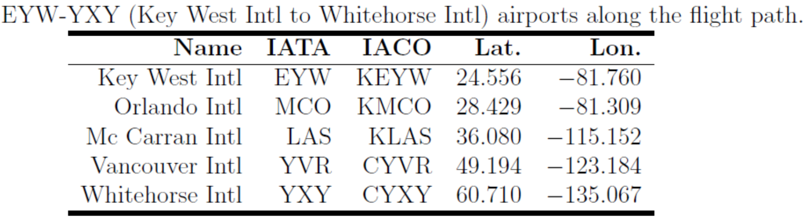
Choosing a source and terminating IATA airport.

Steps to identify the International Air Transport Association (IATA) code for exploration:

1. Search the airport location database for airports of interest.
2. Extract the IATA codes for those airports.
3. Give those codes to the airlinePaths.R program.
4. Evaluate results. Results include:
   - 4.a All airports from the source to the terminating airport.
   - 4.b The path from source to terminating on four different scaled maps.
   - 4.c Automatic inclusion of results into the next printing of the final report.

ICAO stands for International Civil Aviation Organization, a UN organization.[3]

Sample tabular data from EYW-YXY (Key West Intl to Whitehorse Intl).



EYW-YXY (Key West Intl to Whitehorse Intl) airports along the flight path.

Name	IATA	IACO	Lat.	Lon.
Key West Intl	EYW	KEYW	24.556	−81.760
Orlando Intl	MCO	KMCO	28.429	−81.309
Mc Carran Intl	LAS	KLAS	36.080	−115.152
Vancouver Intl	YVR	CYVR	49.194	−123.184
Whitehorse Intl	YXY	CYXY	60.710	−135.067

Figure: Tabular data from EYW-YXY (Key West Intl to Whitehorse Intl).

Sample map data from EYW-YXY (1 of 2).

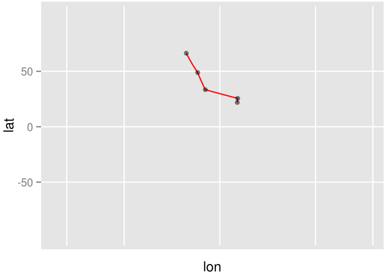


Figure: World view.

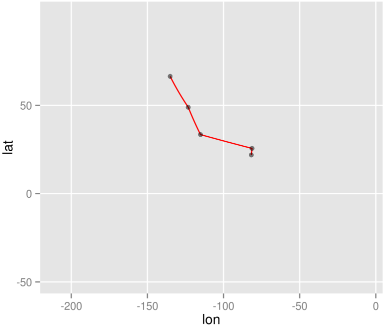


Figure: Hemisphere view.

Sample data from EYW-YXY (2 of 2).

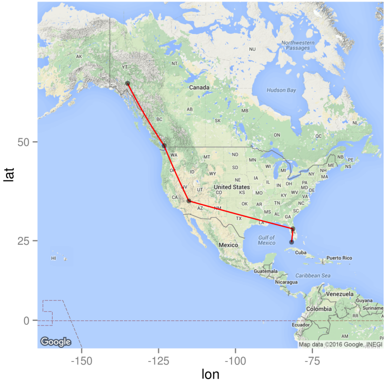


Figure: Continental view

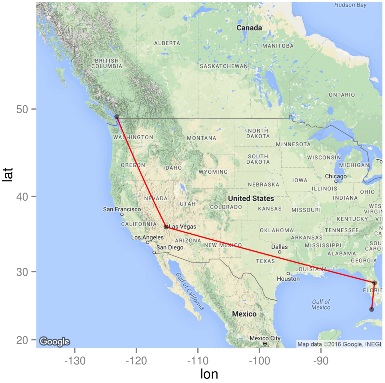
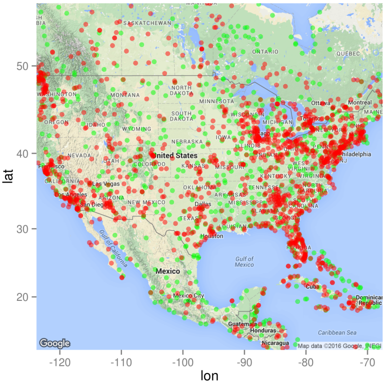
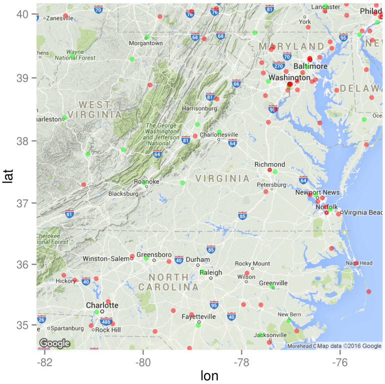


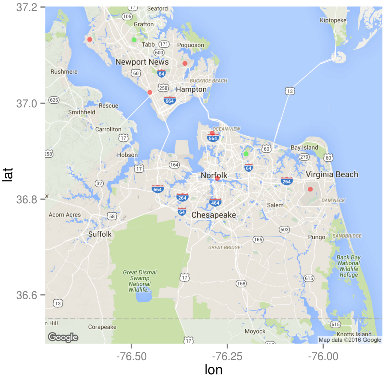
Figure: “Local”

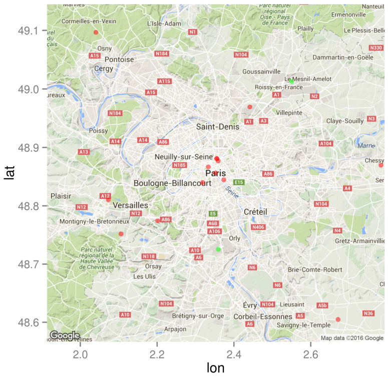
Places that were fun to think about. (1 of 2)





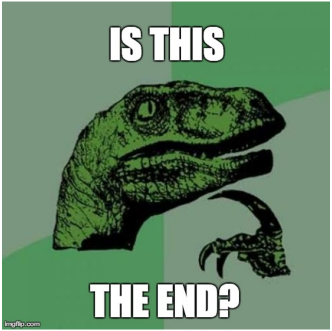
Places that were fun to think about. (2 of 2)





What have we covered?

- Neo4j is a graph database management system optimized to answer questions that can be framed as a graph.
- Neo4j is Java based, with all the strengths and limitations inherent with a Java Virtual Machine.
- The neo4j community edition has practical limitations on the size of the databases it can support.
- Neo4j is very fast and efficient at answering specific types of questions.



Next time: who knows? Columnar databases?

References I

[1] Teo Paoletti, Leonard euler's solution to the königsberg bridge problem, Loci **3** (2011).

[2] BBC Staff, Maths, basic skills, http://www.bbc.co.uk/schools/gcsebitesize/maths/algebra/graphsrev5.shtml, 2016.

[3] IACO Staff, About icao, http://www.icao.int/about-icao/Pages/default.aspx, 2016.