Using Big Data Tools when there are Holes in the Data

Chuck Cartledge

July 26, 2016 at 9:34pm

Table of contents I

Google's BigTable Technology[2]

BigTable characteristics

- Row-keys are printable strings
- Column keys are printable strings
- Qualifiers are arbitrary strings
- Automatic versioning
- Automatic timestamping
- Automatic garbage collection

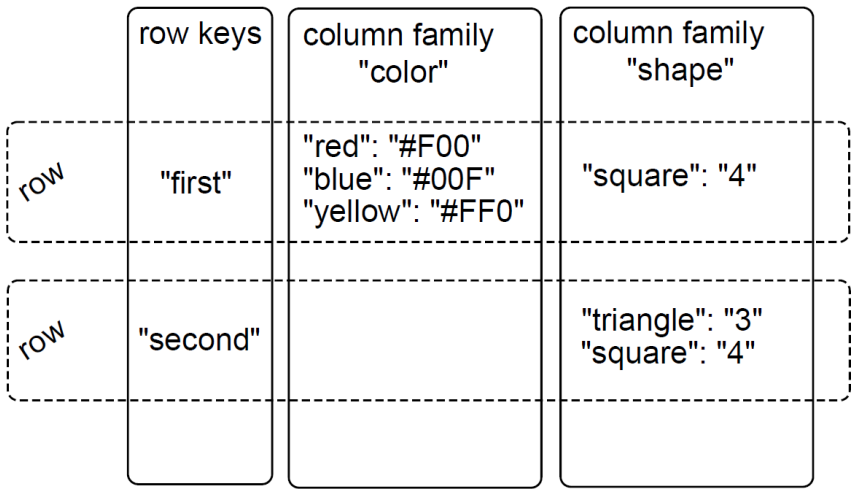


Image from [3].

HBase is a column-oriented database.

HBase ecosystem

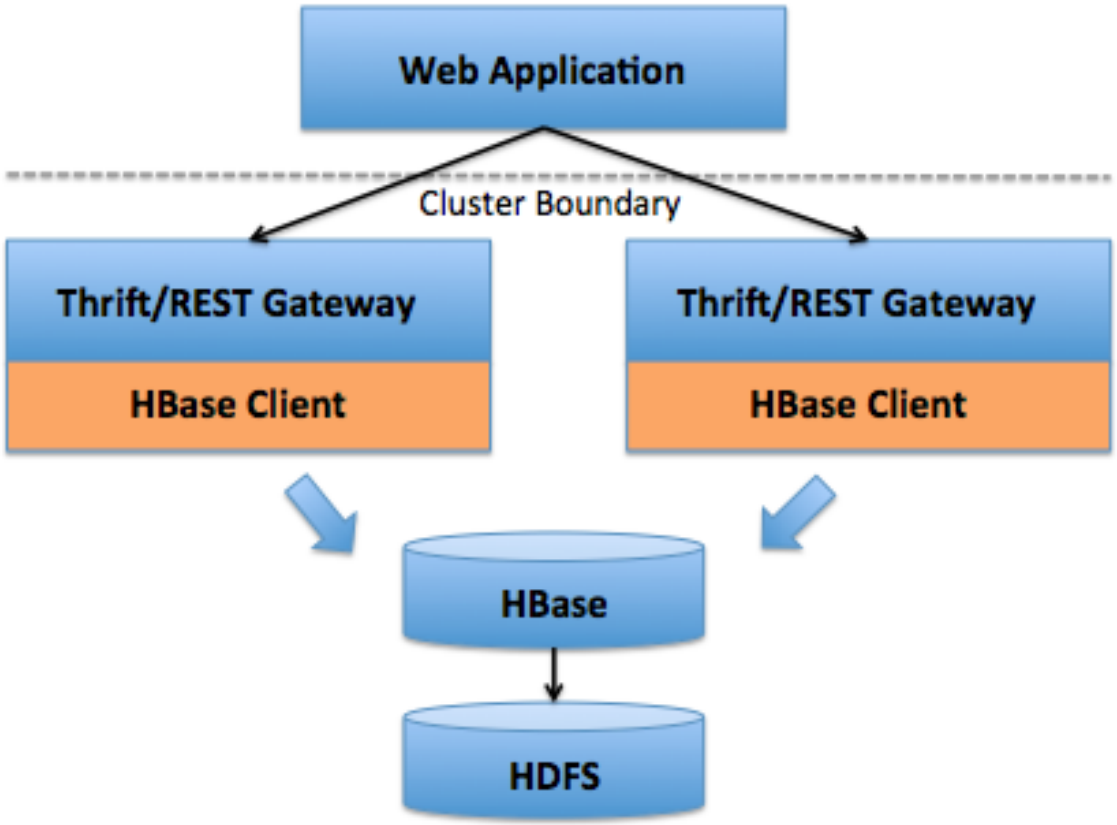
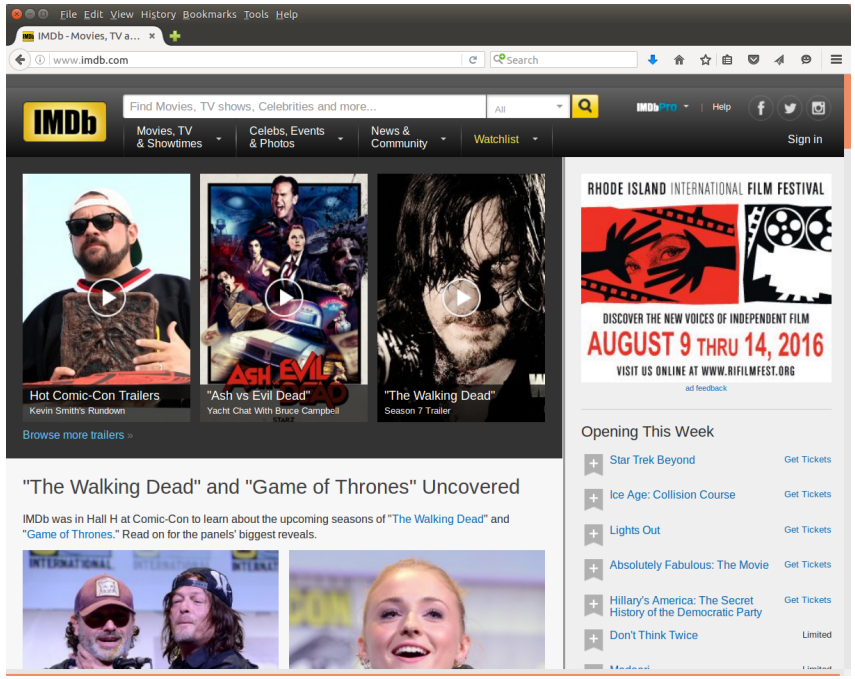


Image from [1].

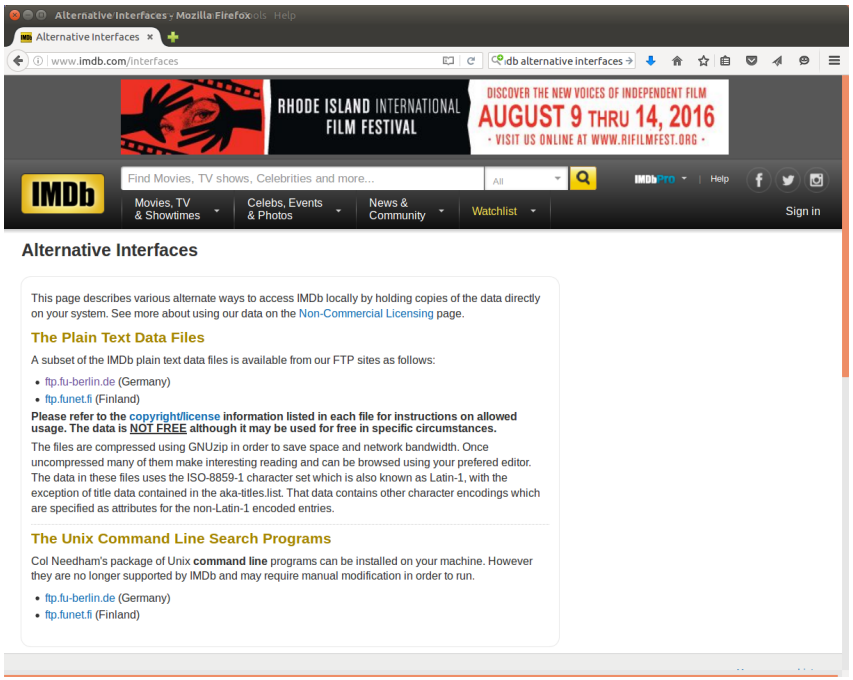
Internet Movie Database (IMDb) goals

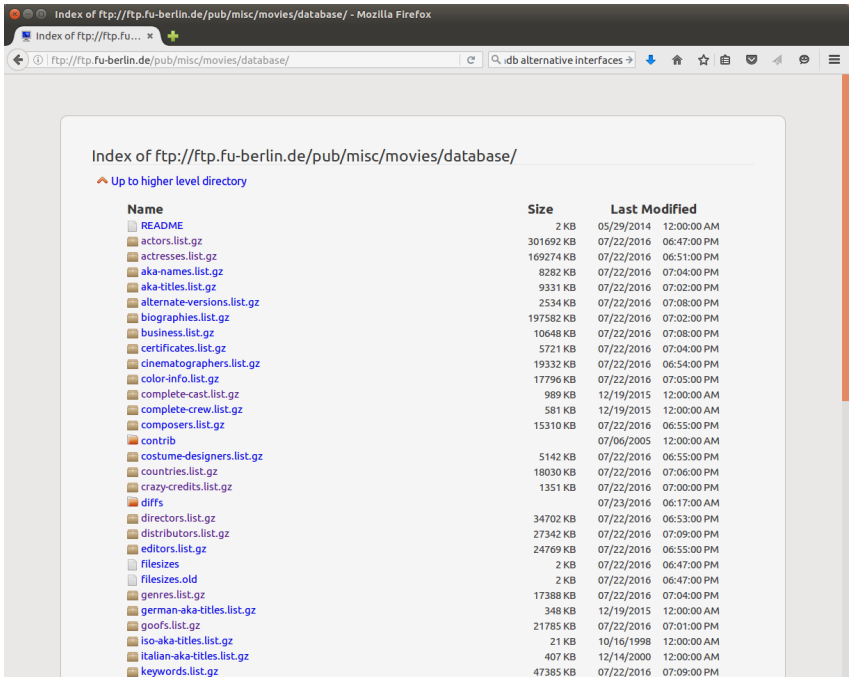
- Number of "crazy credits" per movie
- Number of directors per movie



Use HBase as a database for data from the IMDb.

Where to get the data





`http://www.imdb.com/interfaces`

`ftp://ftp.fu-berlin.de/pub/misc/movies/database/`

Movie data

```
MOVIES LIST
==========

"!Next?" (1994)                                                    1994-1995
"#1 Single" (2006)                                                 2006-????
"#1 Single" (2006) {Cats and Dogs (#1.4)}                          2006
"#1 Single" (2006) {Finishing a Chapter (#1.5)}                    2006
"#1 Single" (2006) {Is the Grass Greener? (#1.1)}                  2006
"#1 Single" (2006) {Stay (#1.8)}                                   2006
"#1 Single" (2006) {The Rules of Dating (#1.3)}                    2006
"#1 Single" (2006) {Timing Is Everything (#1.7)}                   2006
"#1 Single" (2006) {Window Shopping (#1.2)}                        2006
"#1 Single" (2006) {Wingman (#1.6)}                                2006
"#1MinuteNightmare" (2014)                                         2014-????
"#1MinuteNightmare" (2014) {My Sweet Valentine (#1.5)}             ????
"#30Nods" (2014)                                                   2014-2015
"#7DaysLater" (2013)                                               2013-????
"#7DaysLater" (2013) {Apocalypse (#1.2)}                           2013
"#7DaysLater" (2013) {Cowboys (#1.5)}                              2013
"#7DaysLater" (2013) {Drama Queen (#1.1)}                          2013
"#7DaysLater" (2013) {Haunted House (#1.6)}                        2013
"#7DaysLater" (2013) {Portrait (#1.4)}                             2013
"#7DaysLater" (2013) {Zombies (#1.3)}                              2013
"#Adulthood" (????)                                                ????
"#ATown" (2014)                                                    2014-????
"#ATown" (2014) {Best Friends Day (#1.10)}                         2014
"#ATown" (2014) {Chicks in Pink, Vomit in a Sink (#1.6)}                2014
"#ATown" (2014) {Dunzo (#1.9)}                                     2014
"#ATown" (2014) {IMPROVments (#1.4)}                               2014
"#ATown" (2014) {Jobs & Juice (#1.2)}                              2014
"#ATown" (2014) {Kayaking Adventure (#1.5)}                        2014
"#ATown" (2014) {Pilot (#1.1)}                                     2014
"#ATown" (2014) {So Fucked Up (#1.7)}                              2014
"#ATown" (2014) {The Breakup Party (#1.8)}                         2014
"#ATown" (2014) {The Greenbelt (#1.3)}                             2014
"#AwkwardMornings" (2014)                                          2014-????
"#AwkwardMornings" (2014) {Best Friends (#1.4)}                    2014
"#AwkwardMornings" (2014) {Boyfriend (#1.2)}                       2014
```

movies.list.gz

Crazy credits

```
CRAZY CREDITS
=============
# "'Allo 'Allo!" (1982) {Prisoners of War (#4.1)}
- Opening credits prologue: STALAG LUFT IV NORMANDY, FRANCE 1941

# "'Orrible" (2001)
- Episode 1.4 ("May the Best Man Win") uses the Buzzcocks' "Ever Fallen In
  Love" as its end theme.
- Episode 1.8 ("New Best Friend") features Johnny Vaughan and Ricky Grover
  singing "Up Where We Belong" as its end theme.

# "1 ret og 2 vrang" (1969) {(1970-03-16)}
- This episode was broadcast with the end credit roll missing.

# "12 oz. Mouse" (2005) {Auraphull (#2.6)}
- This episode was written by "No One."

# "2 Stupid Dogs" (1993) {A Quarter/Egg/Red (#1.6)}
- John Kricfalusi was credited for supplying "Tidbits of Poor Taste."

# "2 Stupid Dogs" (1993) {Family Values/Platypus/Red Strikes Back (#1.10)}
- John Kricfalusi was credited for supplying "Tidbits of Poor Taste."

# "2 Stupid Dogs" (1993) {Stunt Dogs/Doctor O/Return of Red (#1.11)}
- John Kricfalusi was credited for supplying "Tidbits of Poor Taste."

# "21 Jump Street" (1987) {Back from the Future (#4.15)}
- At the end of the episode, instead of the normal credits showing various
  high schools, we are treated to outtakes from over the series.

# "21 Jump Street" (1987) {Don't Stretch the Rainbow (#2.7)}
- qv##nm0455052##'s "I Have a Dream..." speech is played over the closing credits in pla

# "21 Jump Street" (1987) {Gotta Finish the Riff (#1.6)}
- This is the only time in the show's history where the episode title is shown prior to

# "24" (2001)
- Each episode of the show opens with a title screen and Kiefer  Sutherland's
  voice-over saying "The following takes place between (hour) and  (hour)"
```

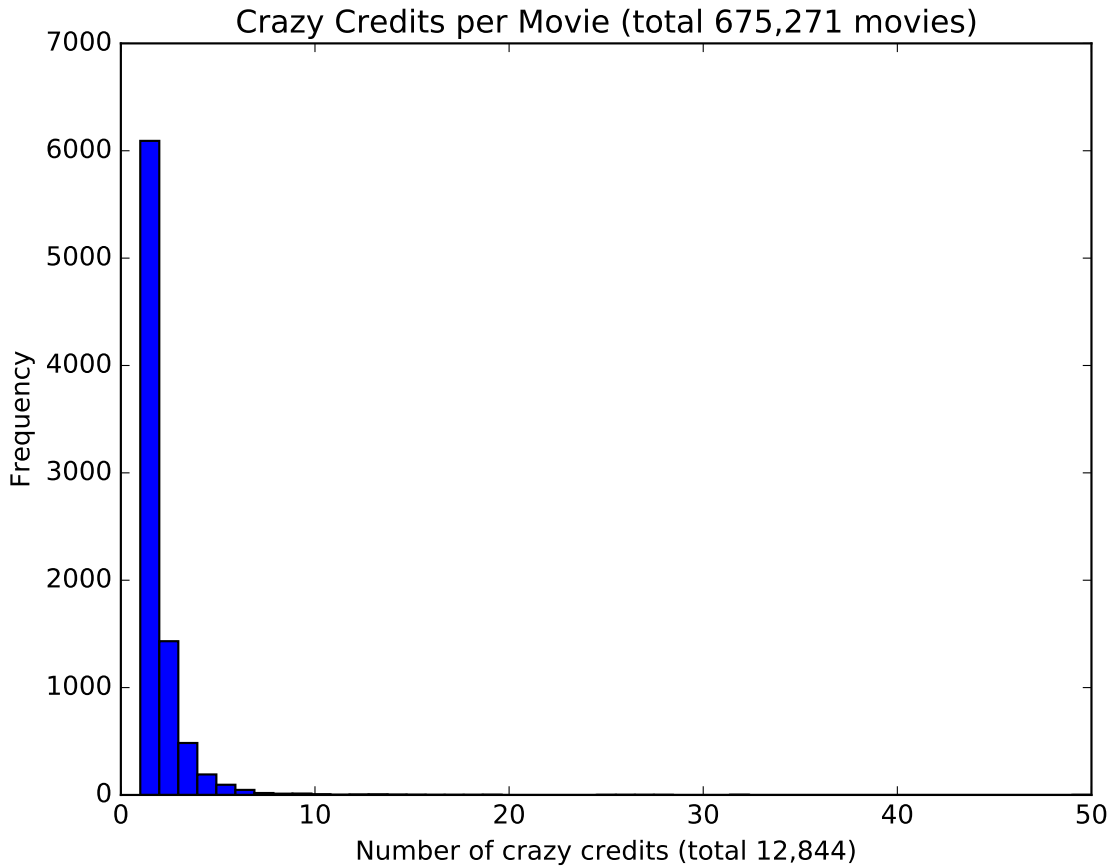
crazy-credits.list.gz

Directors

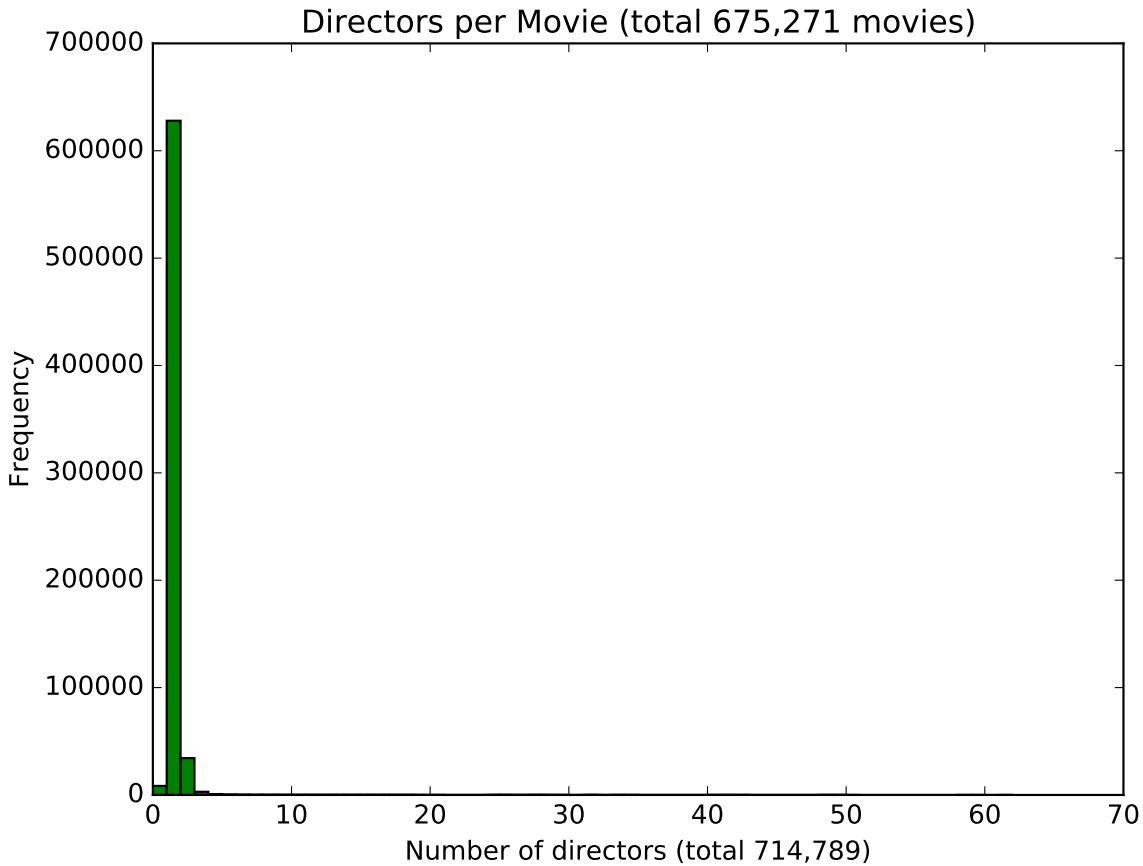
```
------------------------------------------------------------------------------
KEY:

"xxxxx"        = a television series
"xxxxx" (mini) = a television mini-series
(TV)           = TV movie, or made for cable movie
(V)            = made for video movie (this category does NOT include TV
                 episodes repackaged for video, guest appearances in
                 variety/comedy specials released on video, or
                 self-help/physical fitness videos)
(VG)           = videro game


THE DIRECTORS LIST
==================

Name                          Titles
----                          ------
&Oumlzkul, Ahmet Salih        Ii (2013)

'Abd Al-Hamid, Ja'far         A Two Hour Delay (2001)
                              Badgeless sur la Croisette (2012)
                              Just Outside the Frame: The Profilmic Event and Beyond (2008)
                              Mesocafe (2009) {{SUSPENDED}}
                              Mesocafé (2011)

'D.J'Arlia, Domenic           She'll Never Know (2012)

'Dada' Pecori, Diego          Adam (????)  (attached)
                              Cantarella (2011)
                              Makhno Beer (2010)

'Kid Niagara' Kallet, Harry   Drug Demon Romance (2012)  (co-director)

'Kusare, Mak (I)              Baby Beautiful (2013/II)
                              Comrade (2008)

'Kusare, Mak (II)             A Play Called a Temple Made of Clay (2014)

'Legend' Spivey, Larry        The Crime City Diaries: Entry 1 - Crooked (2012)

'Noble Julz'Hamilton, Ulia    Church Hurt (2015)

's Gravesande, Ad             "Het gat van Nederland" (1972)

's-Gravesande, Pjotr          Editors: The Back Room (Festival Edition) (2005) (V)
                              Go Back to the Zoo: Live at Paradiso 2011 (2011) (V)
                              Lucie Silvas: Live in Amsterdam (2007) (V)
                              Milow: Maybe Next Year - Live in Amsterdam (2009) (V)  (uncredite
d)
```

directors.list.gz

Credit results

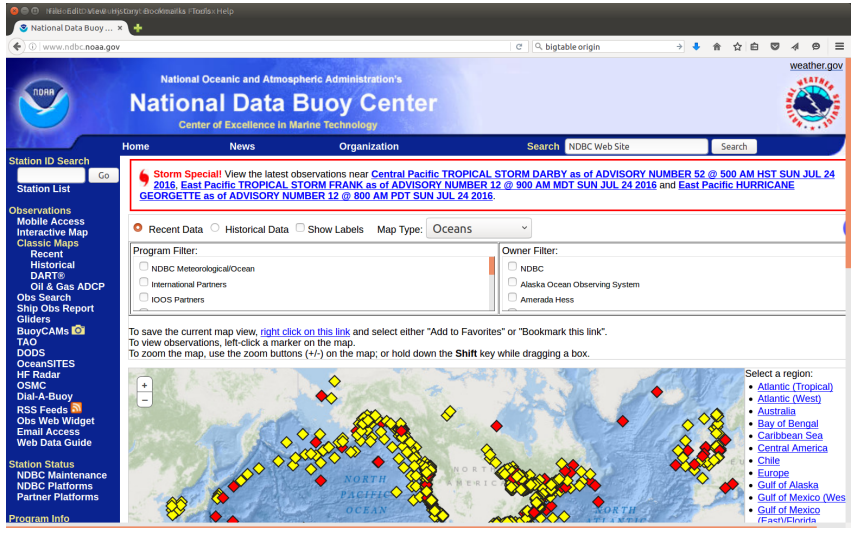


Director results



National Data Buoy Center (NDBC) goals

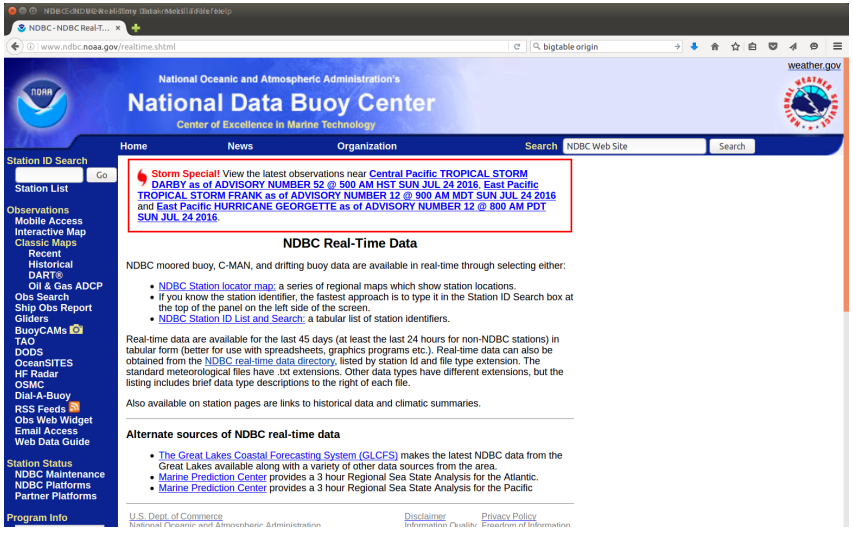
- Explore "real-time data"
- Demonstrate interfacing HBase with dynamic data



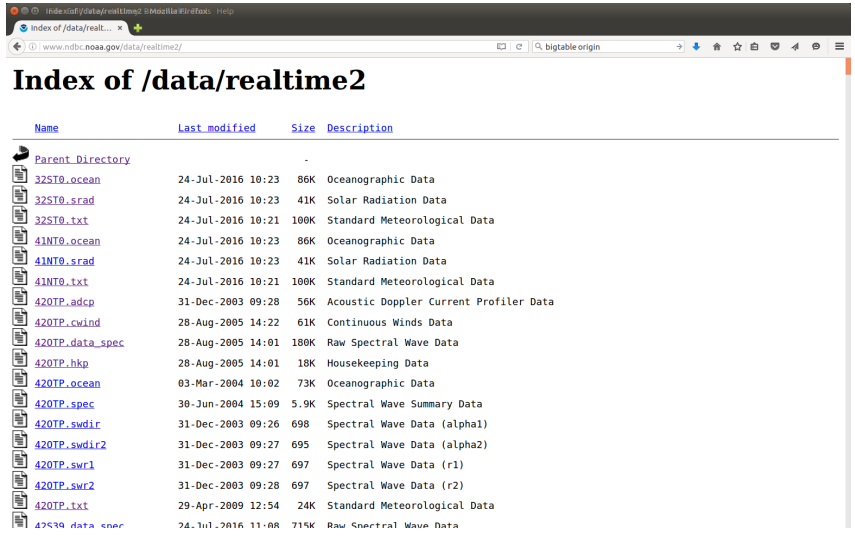
`http://www.ndbc.noaa.gov/`

Explore sources of data used to make weather predictions.

Where to get the data



http://www.ndbc.noaa.gov/realtime.shtml



http://www.ndbc.noaa.gov/data/realtime2/

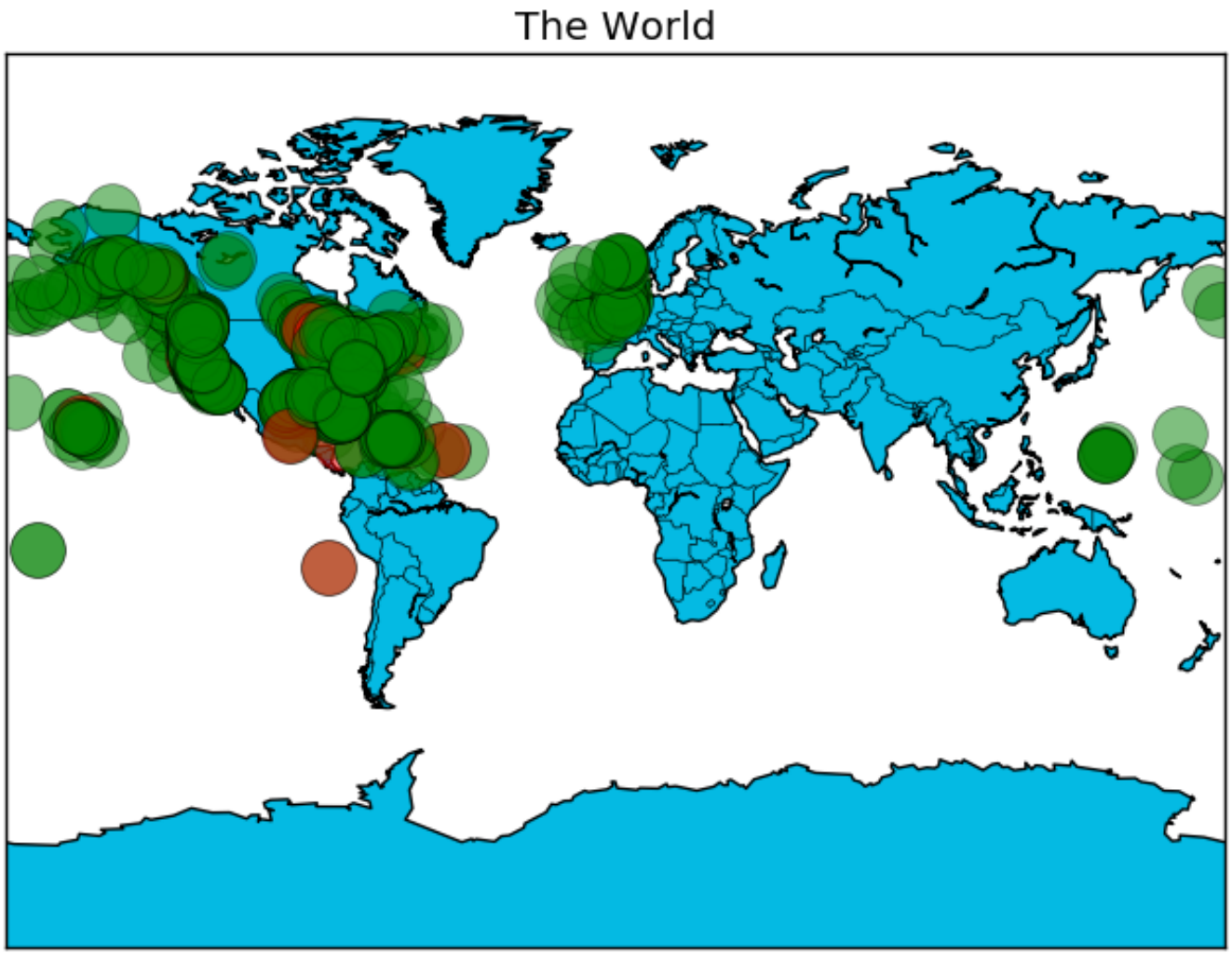
Sample data from station 41013

#YY	MM	DD	hh	mm	WDIR	WSPD	GST	WVHT	DPD	APD	MWD	PRES	ATMP	WTMP	DEWP	VIS	PTDY	TIDE
#yr	mo	dy	hr	mn	degT	m/s	m/s	m	sec	sec	degT	hPa	degC	degC	degC	nmi	hPa	ft
2016	07	22	17	30	130	1.0	2.0	MM	MM	MM	MM	1019.5	28.6	29.9	21.7	MM	MM	MM
2016	07	22	17	20	120	1.0	2.0	MM	MM	MM	MM	1019.6	28.7	29.9	22.0	MM	MM	MM
2016	07	22	17	10	80	1.0	2.0	MM	MM	MM	MM	1019.8	28.5	29.9	21.7	MM	MM	MM
2016	07	22	17	00	100	1.0	3.0	MM	MM	MM	MM	1019.6	28.6	29.8	21.9	MM	+0.5	MM
2016	07	22	16	50	80	1.0	2.0	0.8	9	6.8	116	1019.5	28.6	29.8	21.9	MM	MM	MM
2016	07	22	16	40	90	1.0	3.0	MM	MM	MM	MM	1019.6	28.6	29.7	21.7	MM	MM	MM
2016	07	22	16	30	80	2.0	3.0	MM	MM	MM	MM	1019.5	28.4	29.7	21.4	MM	MM	MM
2016	07	22	16	20	70	2.0	3.0	MM	MM	MM	MM	1019.6	28.4	29.7	21.8	MM	MM	MM
2016	07	22	16	10	70	2.0	3.0	MM	MM	MM	MM	1019.5	28.4	29.7	21.7	MM	MM	MM
2016	07	22	16	00	50	2.0	3.0	MM	MM	MM	MM	1019.5	28.4	29.7	21.5	MM	+0.7	MM
2016	07	22	15	50	50	2.0	3.0	0.8	10	6.9	117	1019.5	28.4	29.7	21.5	MM	MM	MM
2016	07	22	15	40	40	1.0	2.0	MM	MM	MM	MM	1019.5	28.4	29.7	21.3	MM	MM	MM
2016	07	22	15	30	10	1.0	2.0	MM	MM	MM	MM	1019.8	28.6	29.6	21.0	MM	MM	MM
2016	07	22	15	20	30	1.0	2.0	MM	MM	MM	MM	1020.0	28.5	29.6	21.1	MM	MM	MM
2016	07	22	15	10	40	2.0	3.0	MM	MM	MM	MM	1019.9	28.4	29.6	21.1	MM	MM	MM
2016	07	22	15	00	40	3.0	4.0	MM	MM	MM	MM	1019.6	28.2	29.6	21.1	MM	+0.6	MM
2016	07	22	14	50	40	3.0	4.0	0.8	9	6.8	114	1019.5	28.2	29.6	21.3	MM	MM	MM
2016	07	22	14	40	40	3.0	4.0	MM	MM	MM	MM	1019.3	28.2	29.6	21.5	MM	MM	MM
2016	07	22	14	30	50	3.0	4.0	MM	MM	MM	MM	1019.4	28.2	29.6	21.8	MM	MM	MM
2016	07	22	14	20	40	3.0	4.0	MM	MM	MM	MM	1019.3	28.1	29.6	21.7	MM	MM	MM
2016	07	22	14	10	40	3.0	4.0	MM	MM	MM	MM	1019.1	28.1	29.5	21.7	MM	MM	MM

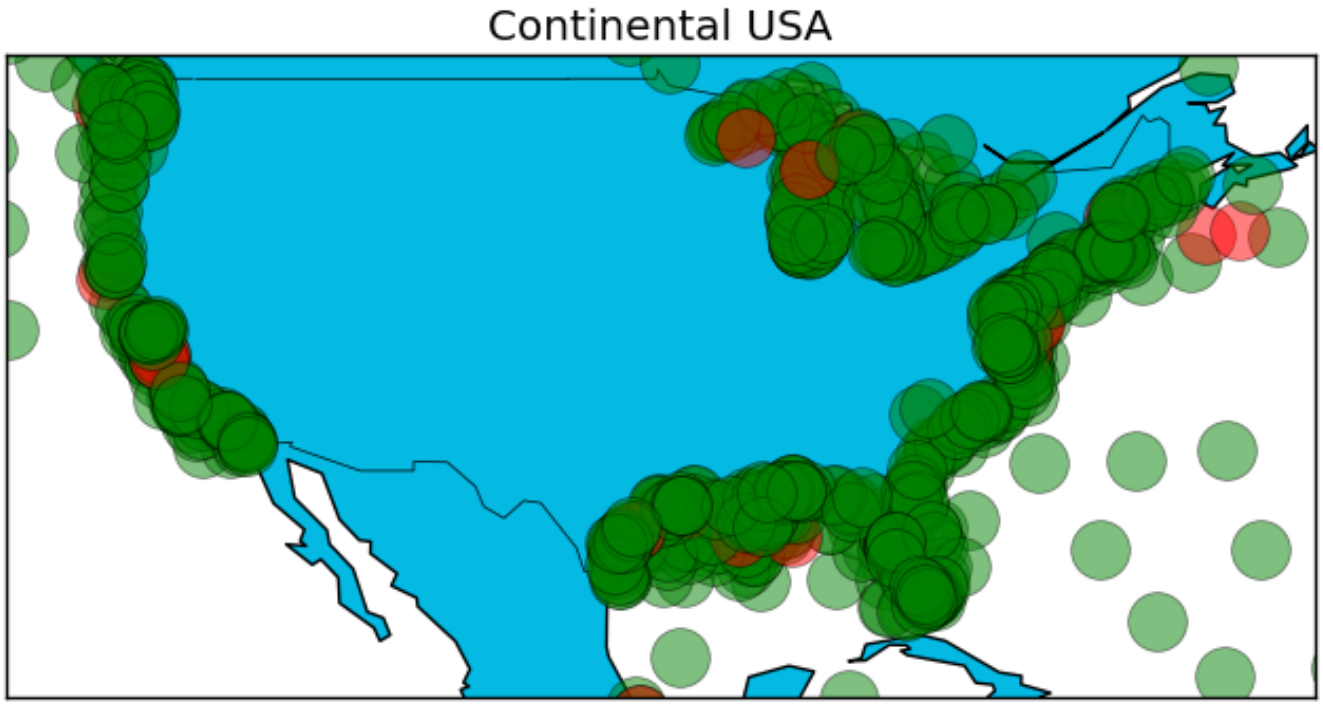
Sample location data for all stations

```
# STATION_ID | OWNER | TTYPE | HULL | NAME | PAYLOAD | LOCATION | TIMEZONE | FORECAST | NOTE
#
00922|DU|Slocum Glider||OTN201 - 4800922||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|E| |
00923|DU|Slocum Glider||OTN200 - 4800923||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|E| |
01500|R|Spray Glider||SP031 - 3801500||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|?| |
01502|UA|Slocum Glider||Penobscot - 4801502||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|E| |
01503|WH|Slocum Glider||Saul - 4801503||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|E| |
01504|UM|Slocum Glider||Blue - 4801504||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|E| |
01505|RU|Slocum Glider||RU28 - 4801505||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|E| |
01506|RU|Slocum Glider||RU22 - 4801506||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|E| |
01507|RU|Slocum Glider||RU23 - 4801507||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|E| |
01508|UD|Slocum Glider||OTIS - 4801508||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|E| |
01510|CS|Slocum Glider||Salacia - 4801510||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|E| |
01511|S|Slocum Glider||Modena - 4801511||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|E| |
01517|WH|Slocum Glider||WHOI_406 - 4801517||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|E| |
01518|RU|Slocum Glider||RU30 - 4801518||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|E| |
01519|UA|Slocum Glider||Unit - 4801519||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|E| |
01521|R|Spray Glider||SP011 - 4801521||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|?| |
01522|R|Spray Glider||SP018 - 4801522||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|?| |
01523|R|Spray Glider||SP025 - 4801523||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|?| |
01524|R|Spray Glider||SP028 - 4801524||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|?| |
01526|R|Spray Glider||SP048 - 4801526||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|?| |
01531|R|Spray Glider||SP407 - 4801531||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|?| |
01532|R|Spray Glider||SP020 - 4801532||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|?| |
01534|R|Spray Glider||SP030 - 4801534||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|?| |
01535|R|Spray Glider||SP052 - 4801535||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|?| |
01536|R|Spray Glider||SP063 - 4801536||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|?| |
01537|RU|Slocum Glider||ru07 - 4801537||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|| |
01538|R|Spray Glider||SP043 - 4801538||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|?| |
01909|R|Spray Glider||SCRIPPS Glider - 4801909||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|?| |
01910|R|Spray Glider||SCRIPPS Glider - 4801910||30.000 N 90.000 W (30&#176;0'0" N 90&#176;0'0" W)|?| |
0y2w3|CG|Weather Station||Sturgeon Bay CG Station, WI||44.794 N 87.313 W (44&#176;47'39" N 87&#176;18'48" W)|C|
13001|PR|Atlas Buoy|PM-595|NE Extension||12.000 N 23.000 W (12&#176;0'0" N 23&#176;0'0" W)|| |
13002|PR|Atlas Buoy||NE Extension||21.000 N 23.000 W (21&#176;0'0" N 23&#176;0'0" W)|| |
13008|PR|Atlas Buoy|PM-531|Reggae||15.000 N 38.000 W (15&#176;0'0" N 38&#176;0'0" W)|| |
13009|PR|Atlas Buoy|PM-533|Lambada||8.000 N 38.000 W (8&#176;0'0" N 38&#176;0'0" W)|| |
13010|PR|Atlas Buoy|PM-590|Soul||0.000 N 0.000 E (0&#176;0'0" N 0&#176;0'0" E)|| |
15001|PR|Atlas Buoy|PM-597|Gavotte||10.000 S 10.000 W (10&#176;0'0" S 10&#176;0'0" W)|| |
15002|PR|Atlas Buoy|PM-591|Java||0.000 N 10.000 W (0&#176;0'0" N 10&#176;0'0" W)|| |
15006|PR|Atlas Buoy|PM-593|Valse||6.000 S 10.000 W (6&#176;0'0" S 10&#176;0'0" W)|| |
15007|PR|Atlas Buoy|PM-605A|SE Extension||6.000 S 8.000 E (6&#176;0'0" S 8&#176;0'0" E)|| |
```

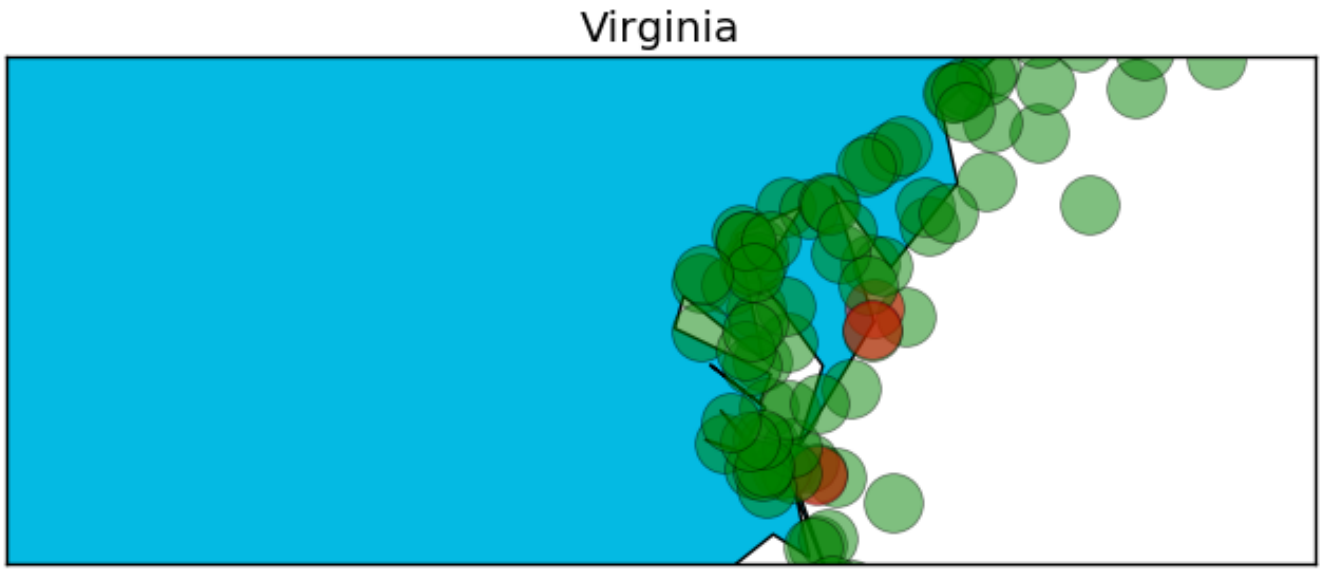
Real time buoy data around the world



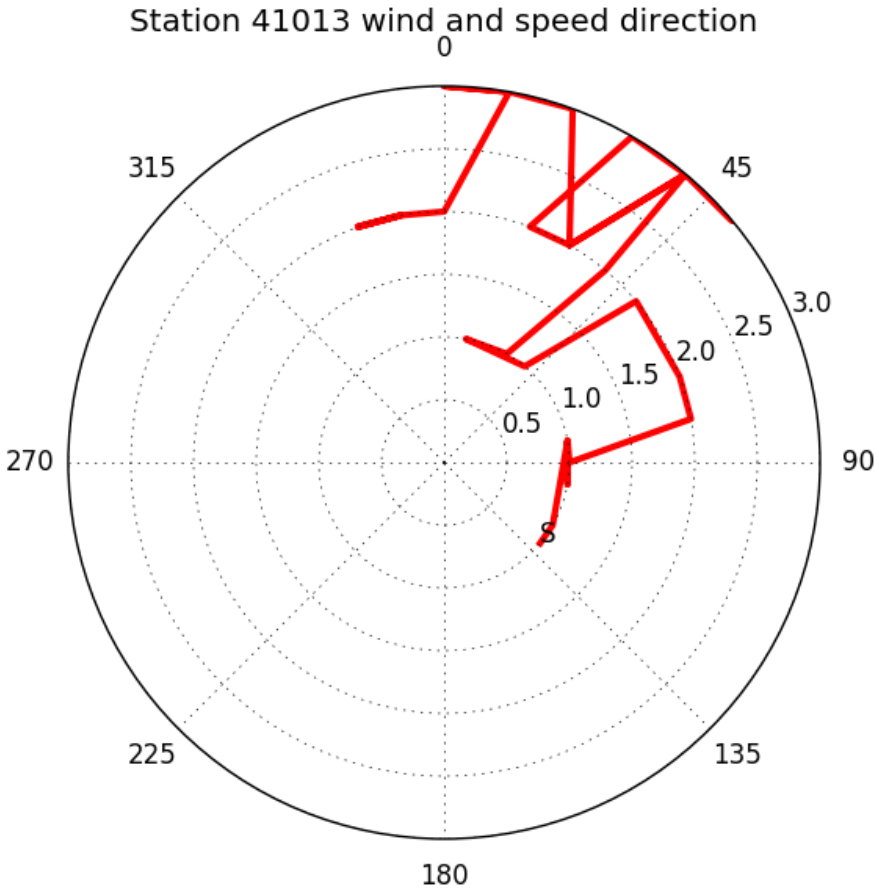
Real time buoy data around the continental US



Real time buoy data around VA

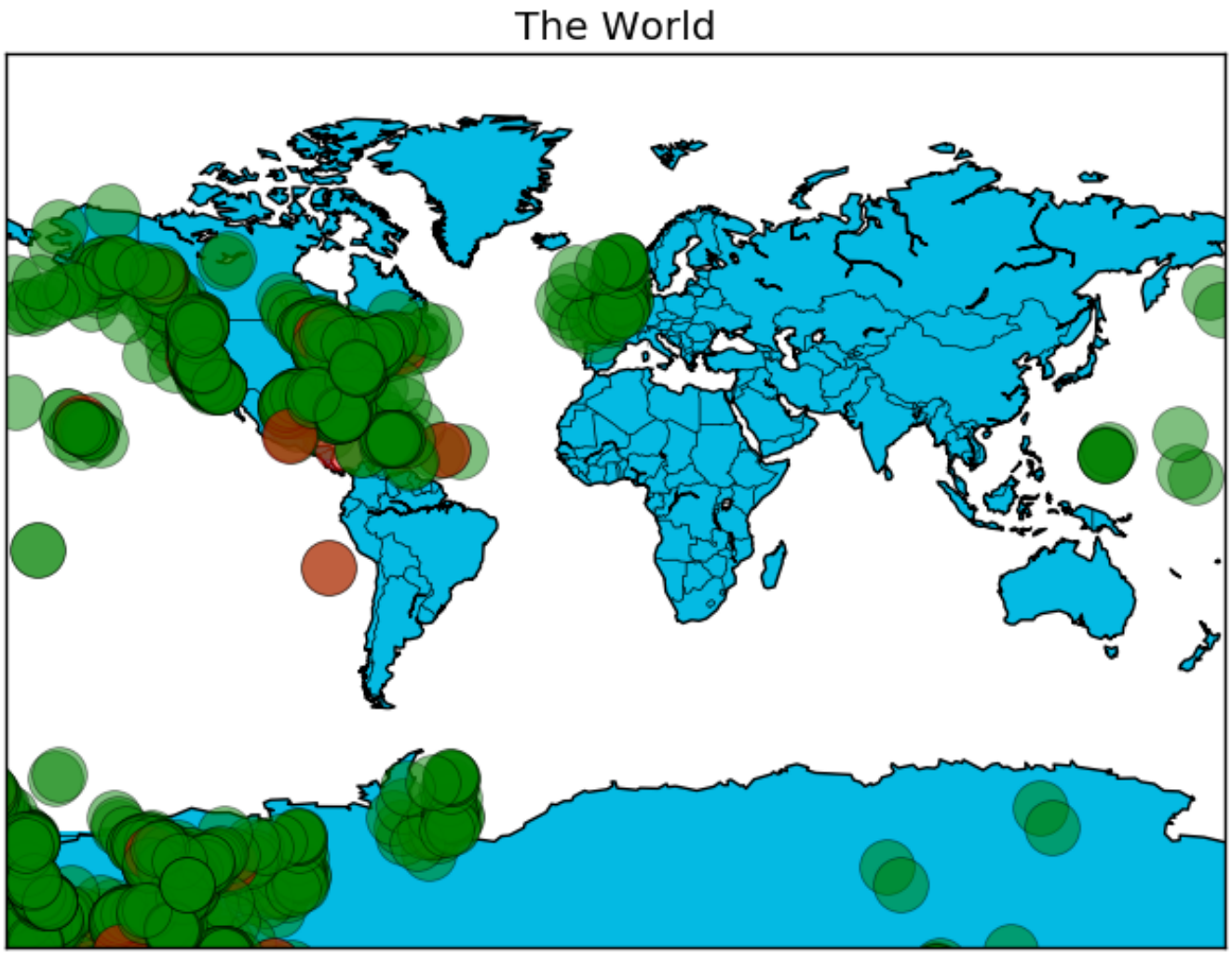


Real time wind data from station 41013



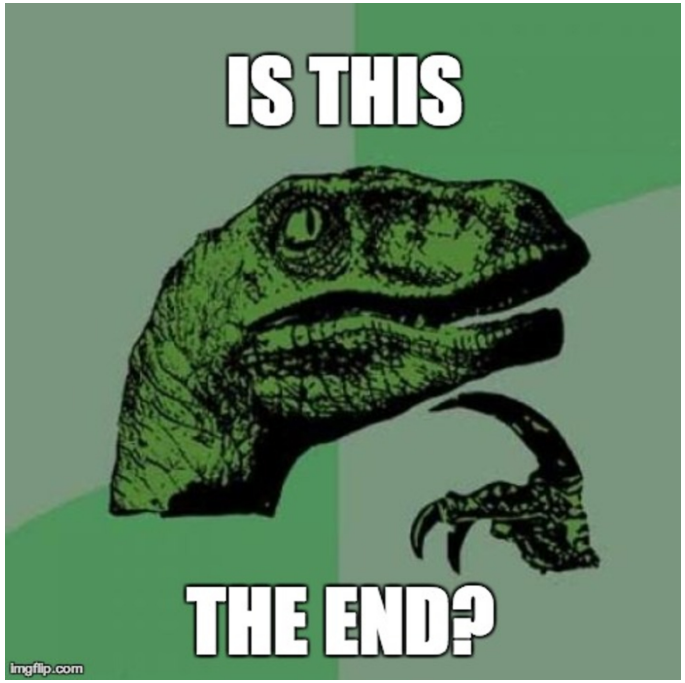
The "S" marks the oldest data.

Interesting things about the Python matplotlib plot



What have we covered?

- HBase ecosystem is Java based, with all the strengths and limitations inherent with a Java Virtual Machine.
- Parsing of static and real time data wasn't too difficult
- HBase is columnar database system that handle "holes" in data well



Next time: who knows? Document databases?

References I

[1] Jesse Anderson, How-to: Use the hbase thrift interface, part 1, http://blog.cloudera.com/blog/2013/09/how-to-use-the-hbase-thrift-interface-part-1/, 2013.

[2] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C Hsieh, Deborah A Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E Gruber, Bigtable: A distributed storage system for structured data, ACM Transactions on Computer Systems (TOCS) **26** (2008), no. 2, 4.

[3] Eric Redmond and Jim R Wilson, Seven databases in seven weeks, Pragmatic Bookshelf, 2012.