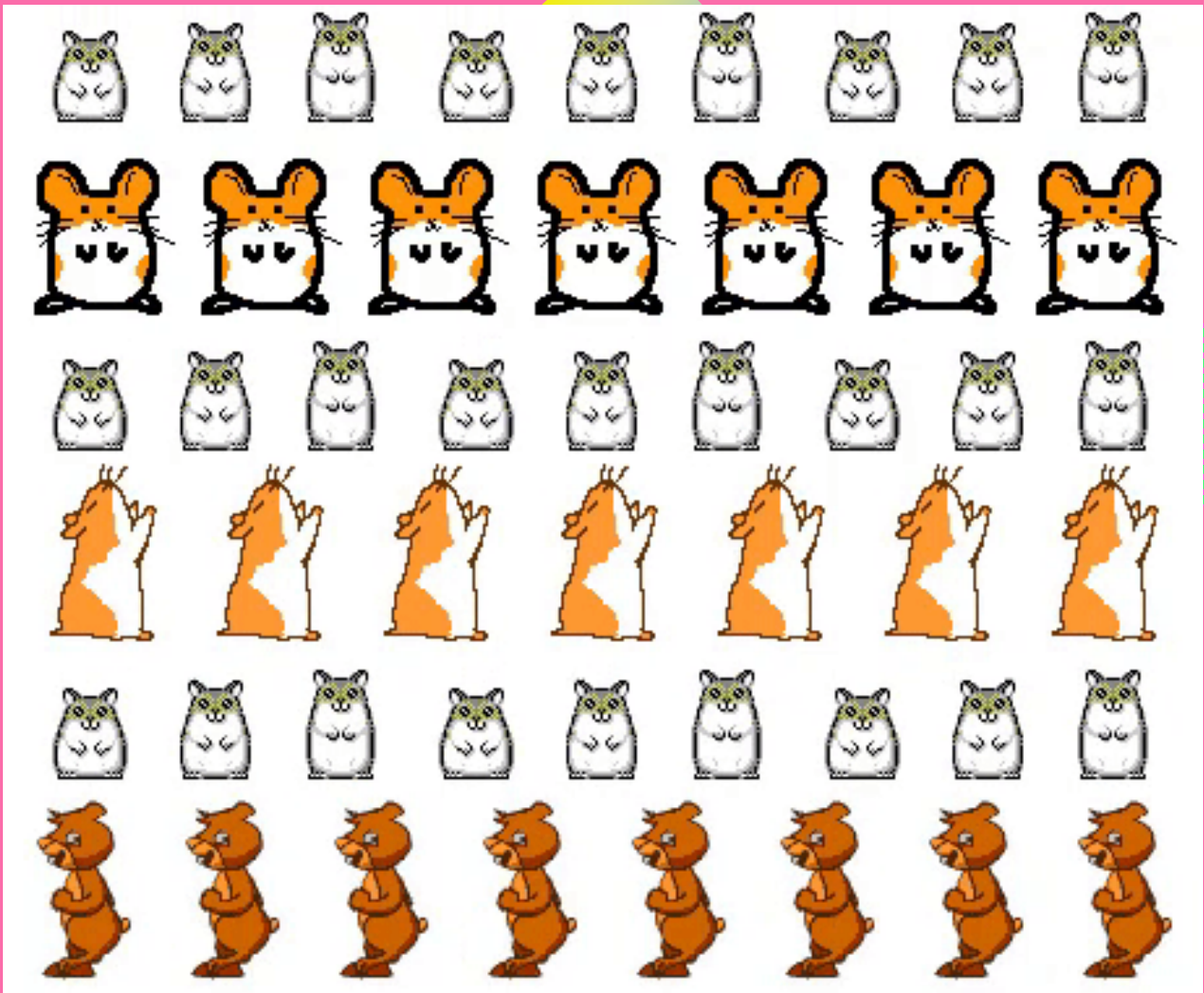
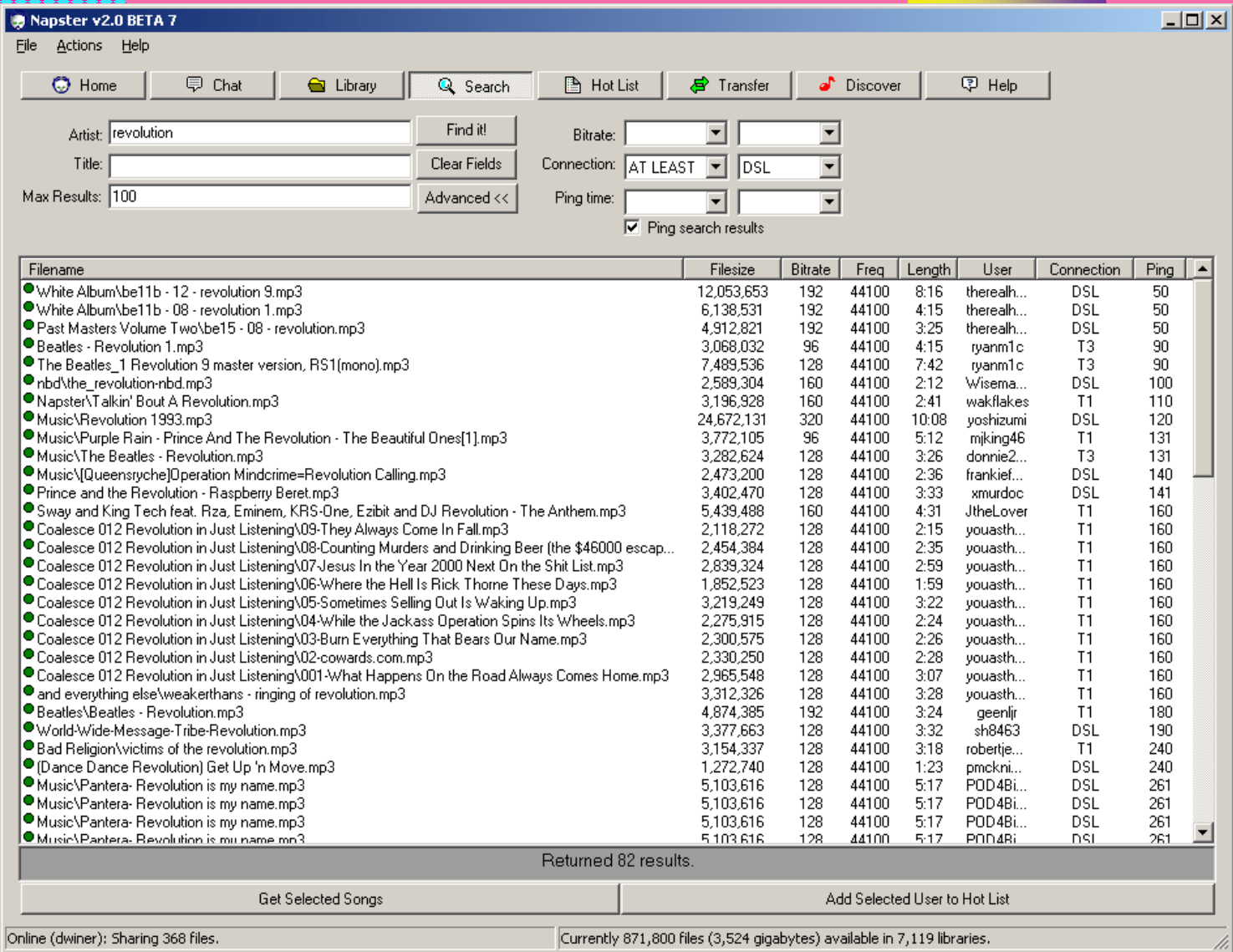



Cross-Server Data Joins on Slow Networks with Python

Bert Wagner

20
YEARS CELEBRATION
PYCON US
23


“MOM, DON’T USE THE PHONE”





It's not news, it's **FARK.com**






Advertise on Fark.com - [Small](#), [Large](#),



Search

☒ Web ☐ Fark.com

To read articles, click the icon left of the entry. Rinse. Repeat. Wipe hands on pants.

# February 09, 2005:	SUBMIT A LINK	# of Comments
<div style="display: flex; align-items: center;"> <div style="margin-right: 10px;"> (Some Guy) </div> <div style="background-color: #f0f0f0; padding: 2px 5px; font-size: 0.8em; margin-right: 10px;">INTERESTING</div> <div>Instead of making taxpayers foot the bill for prison costs, why not let the prisoners pay for it themselves?</div> </div>		(64)
<div style="display: flex; align-items: center;"> <div style="margin-right: 10px;"> (Some Demolition Site) </div> <div style="background-color: #f0f0f0; padding: 2px 5px; font-size: 0.8em; margin-right: 10px;">PHOTOGRAPH</div> <div>Photoshop this building, or how it got this way</div> </div>		(75)
<div style="display: flex; align-items: center;"> <div style="margin-right: 10px;">  </div> <div style="background-color: #f0f0f0; padding: 2px 5px; font-size: 0.8em; margin-right: 10px;">WEIRD</div> <div>Plucked chickens are falling from the sky and damaging roofs in New South Wales. "There's something unusual going on," notes area resident Joe Obvious</div> </div>		(15)
<div style="display: flex; align-items: center;"> <div style="margin-right: 10px;">  </div> <div style="background-color: #f0f0f0; padding: 2px 5px; font-size: 0.8em; margin-right: 10px;">WEIRD</div> <div>A record 2,201 cases of "train molestation" were reported in Tokyo last year</div> </div>		(32)
<div style="display: flex; align-items: center;"> <div style="margin-right: 10px;"> (SILive.com) </div> <div style="background-color: #f0f0f0; padding: 2px 5px; font-size: 0.8em; margin-right: 10px;">OBVIOUS</div> <div>Homeless man freezes to death after seeking shelter in abandoned ice factory</div> </div>		(24)
<div style="display: flex; align-items: center;"> <div style="margin-right: 10px;">  </div> <div style="background-color: #f0f0f0; padding: 2px 5px; font-size: 0.8em; margin-right: 10px;">AMUSING</div> <div>The 10 most disrespected entities in sports today</div> </div>		(40)
<div style="display: flex; align-items: center;"> <div style="margin-right: 10px;">  </div> <div style="background-color: #f0f0f0; padding: 2px 5px; font-size: 0.8em; margin-right: 10px;">UNLIKELY</div> <div>Rolling hoop-snake sightings near Rio Grande. Self-propelled hula mushrooms somehow missed. Badgers refuse to comment after seen just wriggling in place</div> </div>		(29)
<div style="display: flex; align-items: center;"> <div style="margin-right: 10px;"> (nbc4.tv) </div> <div style="background-color: #f0f0f0; padding: 2px 5px; font-size: 0.8em; margin-right: 10px;">STRANGE</div> <div>This week's unknown nail in the body belongs to a man with a pain in the neck</div> </div>		(31)
<div style="display: flex; align-items: center;"> <div style="margin-right: 10px;">  </div> <div style="background-color: #f0f0f0; padding: 2px 5px; font-size: 0.8em; margin-right: 10px;">SAD</div> <div>Doobie Brothers drummer Keith Knudsen dies at 56</div> </div>		(48)
<div style="display: flex; align-items: center;"> <div style="margin-right: 10px;"> (Gallup) </div> <div style="background-color: #f0f0f0; padding: 2px 5px; font-size: 0.8em; margin-right: 10px;">INTERESTING</div> <div>Bush's approval rating hits 57 percent. Farkers' heads asplode</div> </div>		(324)

[HELP](#)
[FORUMS](#)
[CHAT](#)
[LITERATURE](#)
[WEBCAMS](#)
[USERS](#)
[NETWORK](#)
[SUBMIT CONTENT](#)
[HTTP://WWW.NEWGROUNDS.COM](http://www.newgrounds.com)



LATEST 5 SUBMISSIONS

1. Mouse Mover
2. Bob's First Cartoon
3. Incinerator
4. Konoha Talent Show
5. Armed Forces

FEATURED NG MEDIA!



New to the site? Read the [Newgrounds Primer!](#)
[Submit](#) your own original work and you could be [awarded \\$250](#) this month!
 Not a Flash author? You can [win \\$100](#) just by voting on submissions!



ALIEN HOMINID - Play this web version and buy the console version, available on PS2 and Gamecube!



MY GOD, ROBOTS! - The third episode of this rockin' new series!



MINI-PUTT 3 - The best Mini Putt game so far by Psycho Goldfish!



CLAVEMAN EP. 4/5 UNO - Tons of crazy jokes and fun for all. This is really some top-notch stuff!



A NEW BUNNY - This might be the funniest thing I've seen in awhile.



DEMONIC DEFENCE 4 - Defend your castle against demonic forces! I love this genre!




digg

login

latest front page stories


207
diggs



submitted by [owmyshoe](#) 14 hours 19 minutes ago (via <http://www.1up.com/do/feature?...>)


digg it

With E3 '06 coming, and the next gen console wars about to take off, 1up.com takes a closer look at the handheld battle between Nintendo's DS, and Sony's PSP.



[36 comments](#) | [blog this](#) | [email this](#) | category: [gaming](#)


219
diggs



submitted by [digitalgopher](#) 6 hours 16 minutes ago (via <http://www.flickr.com/photos/t...>)


digg it

MAKE Flickr photo pool member TommyBear turned a \$5 Staples Easy Button into a switch for his garage door. Here's how you do it, easy!



[22 comments](#) | [blog this](#) | [email this](#) | category: [mods](#)


387
diggs




submitted by [mastersword](#) 3 hours 52 minutes ago (via <http://gadgets.fosfor.se/the-t...>)

digg it

Keyboards come in many shapes - from the simplest computer grey \$9.99 standard 102-key keyboard to variants that seem to come straight out of a Star Trek episode. Here's a top 10 list of some really cool keyboards.



[23 comments](#) | [blog this](#) | [email this](#) | category: [technology](#)




Broadcast Yourself™


[Sign Up](#) | [My Account](#) | [History](#) | [Help](#) | [Log In](#)

[Videos](#)
[Categories](#)
[Channels](#)
[Community](#)
[Upload Videos](#)


Director Videos




[Swing Blade - Parody...](#)
ButterTV



[Hometown Baghdad...](#)
chattheplanet




[Lolla Lives-Shepard...](#)
lollalives




[Black Anthem](#)
pdrog

[See More Featured Videos](#)




Featured Videos selected by:
[YouTube](#)

[Become a guest editor](#)



[Worms Making Music II](#)
01:29
A solo worm performing a thoughtful little number. No worms were hurt.
Following a gruelling audition process, one solo worm w ([more](#))
From: [ashfordaisyak](#) Views: 858 ★★★★★
More in [Howto & DIY](#)




[Leader](#)
00:26
WHITESTKIDS.COM
sketch by the 'Whitest Kids U Know'
From: [whitestkidsdotcom](#) Views: 12,199 ★★★★★
More in [Comedy](#)


My: [Videos](#) - [Favorites](#) - [Playlists](#) - [Inbox](#) - [Subscriptions](#)


Member Login

YouTube Username:
YouTube Password:
 [Sign Up](#)
[Forgot Username](#) | [Forgot Password](#)

New at YouTube

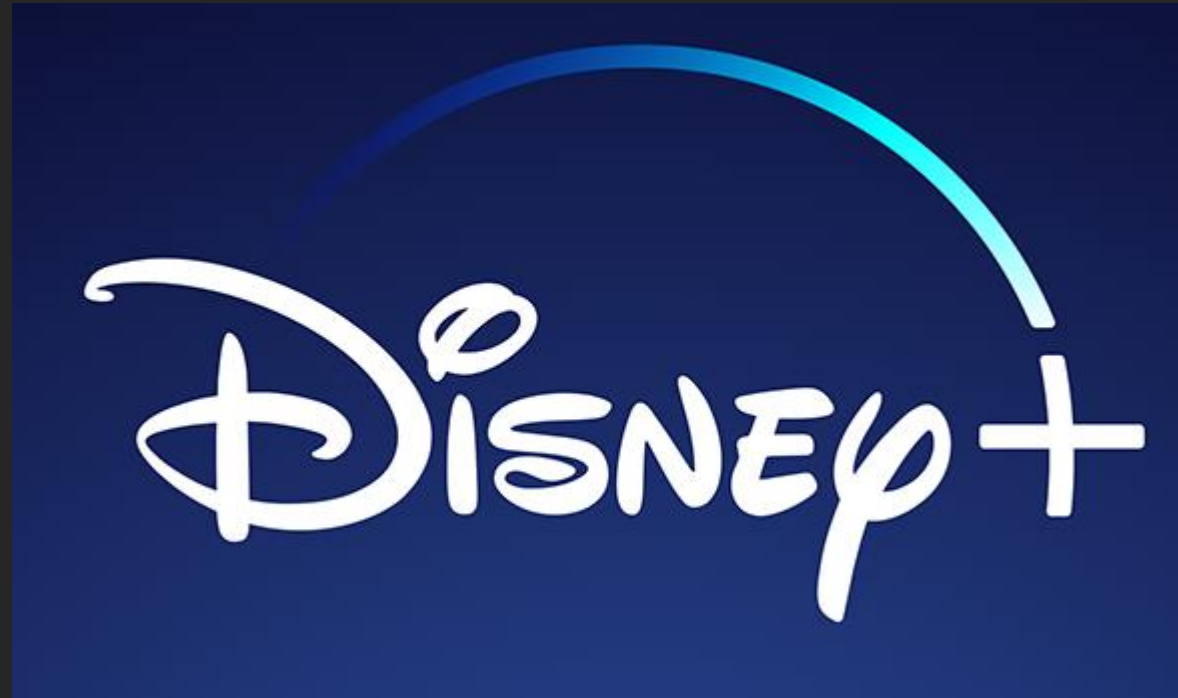

[Channel Customizations](#)
New channel customizations abound! Check [them](#) out in your [channel settings](#).


[Categories](#)
The Categories section has been given a face lift and now shows more featured videos and channels.


[TestTube](#)
More new (and improved) YouTube features in development! Now with Active Sharing.

💀 “STOP STREAMING!” 💀

4



BERT WAGNER

Data Scientist, YouTuber, Blogger

Code and slides:

bertwagner.com/crosstream

Contact:

bertwagner.com

bertwagner@bertwagner.com

[@bertwagner](https://twitter.com/bertwagner)

DATA WITH BERT



AGENDA

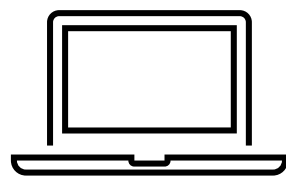
Joining datasets across networks efficiently

- Great Solutions
- Not Bad Solutions
- Ugly Solutions

SCENARIO



CustomerSubset.csv
100k rows
2GB

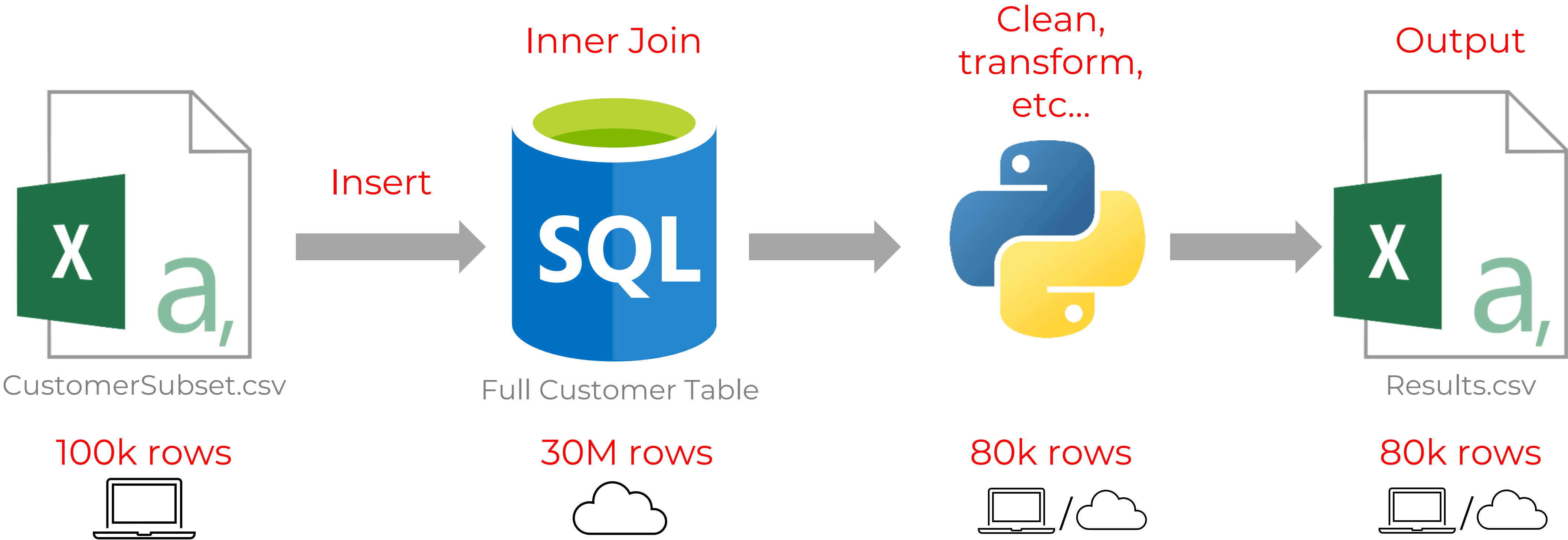


All Customer Data
30M rows
20GB

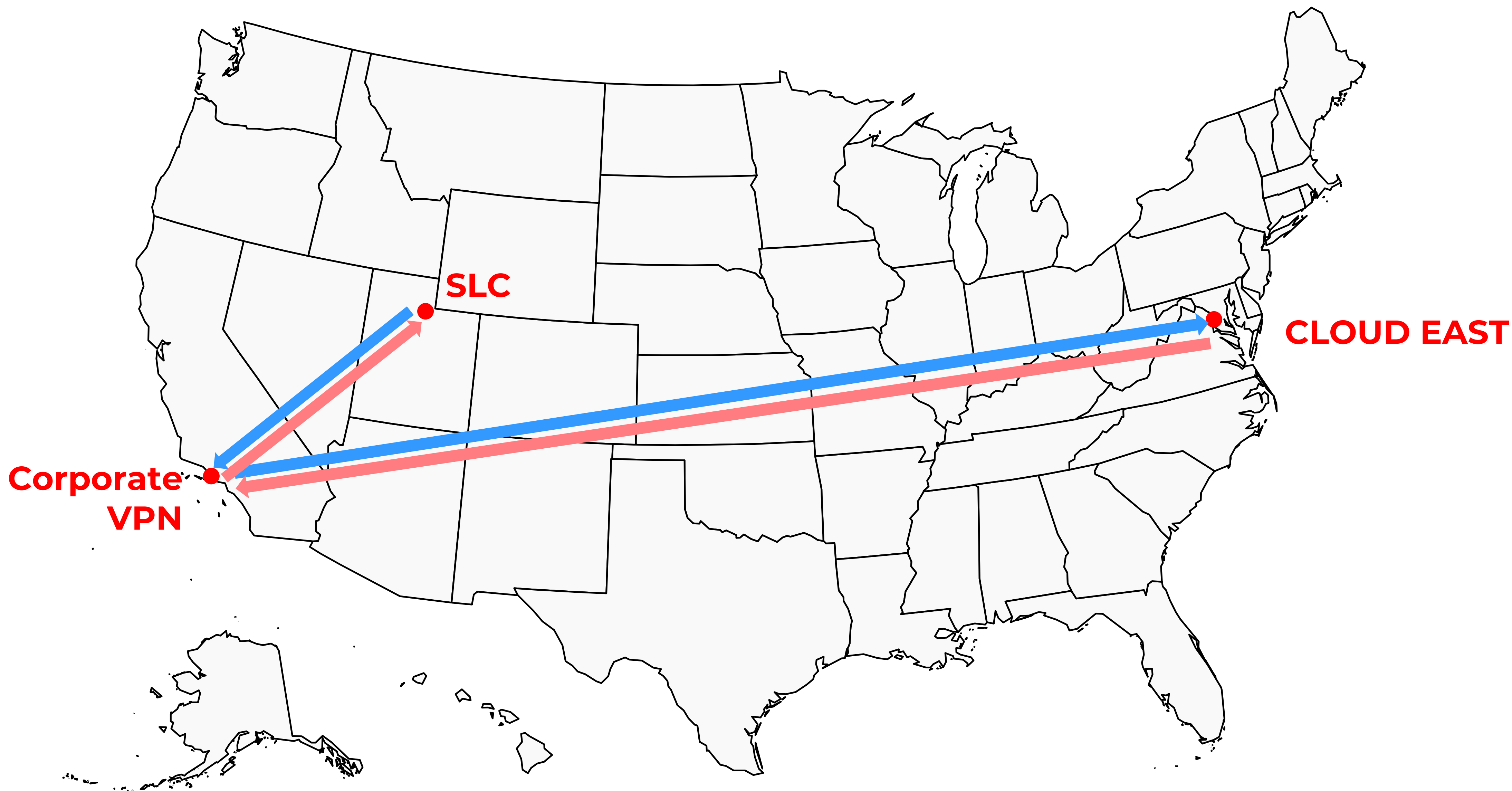


How do we join these datasets together?

SOLUTION: GREAT



SOLUTION: GREAT



SUMMARY: GREAT SOLUTION

- PRO: Minimizes network movement
- PRO: Maximizes using tools for what they are good at
- CON: Might be paying for expensive cloud compute or licensing
- CON: Not good at doing transformations to data before you join

More resources:

- [SQL Anti-Patterns for Analysts](#)
- [Database Indexing for Beginners](#)

SOLUTION: NOT BAD



Full Customer Table



Customer
Subset.csv



Big heavy
duty server



Results.csv



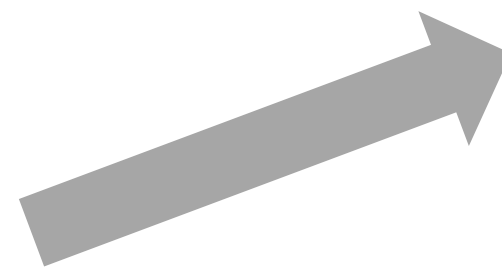
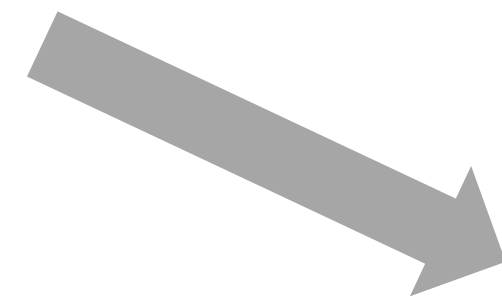
SOLUTION: NOT BAD



Full Customer Table



Customer
Subset.csv



Results.csv



SUMMARY: NOT BAD SOLUTION

- PRO: If large data on same network as servers, performance is good
- PRO: If you have a pool of machines, might be able to parallelize
- PRO: Powerful server will allow you to transform data for your join keys
- CON: You need a big high performance server \$\$\$
- CON: Your server needs to be on the same network as your large dataset

More resources:

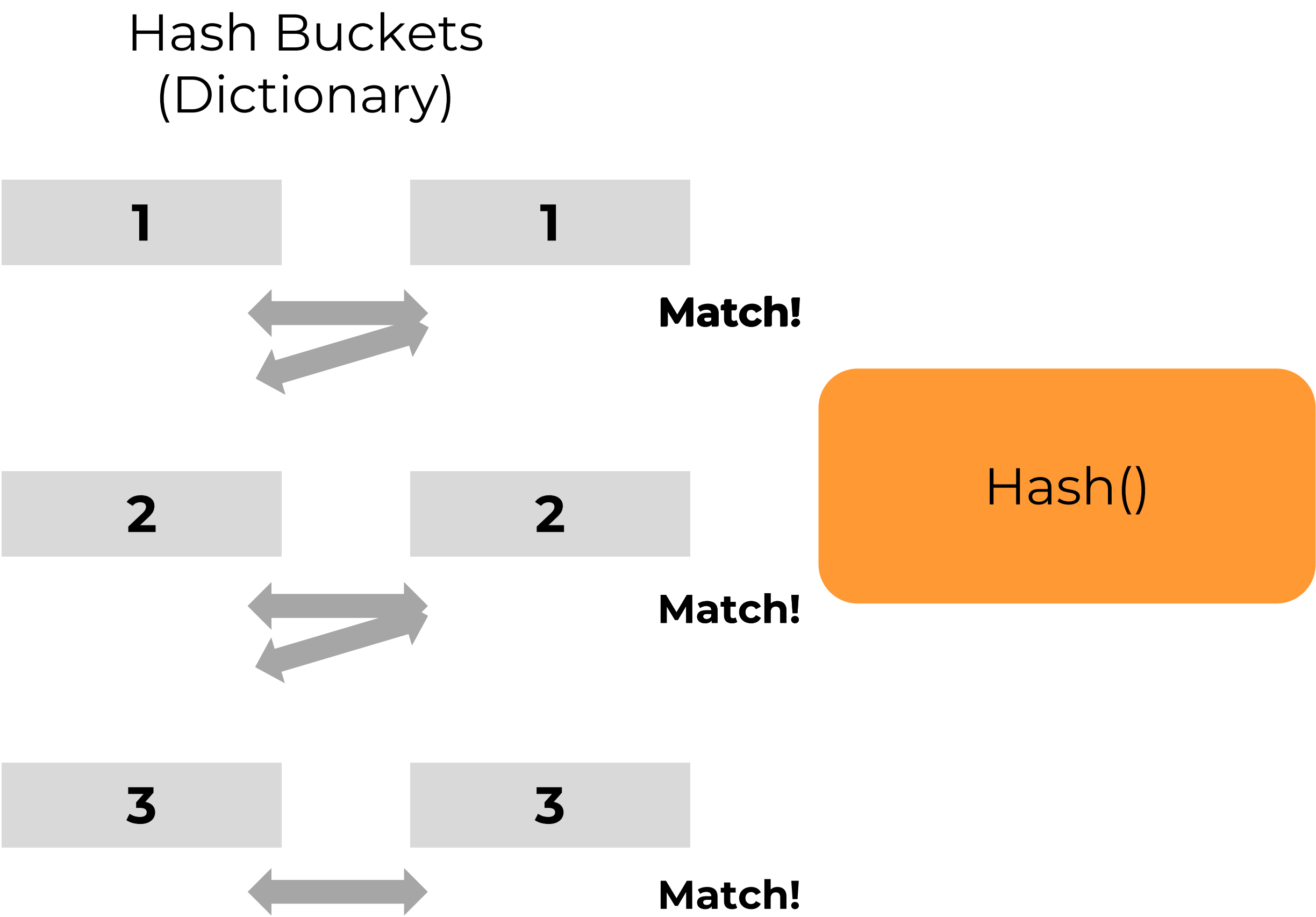
- [Pandas: Scaling to large datasets](#)

SOLUTION: UGLY

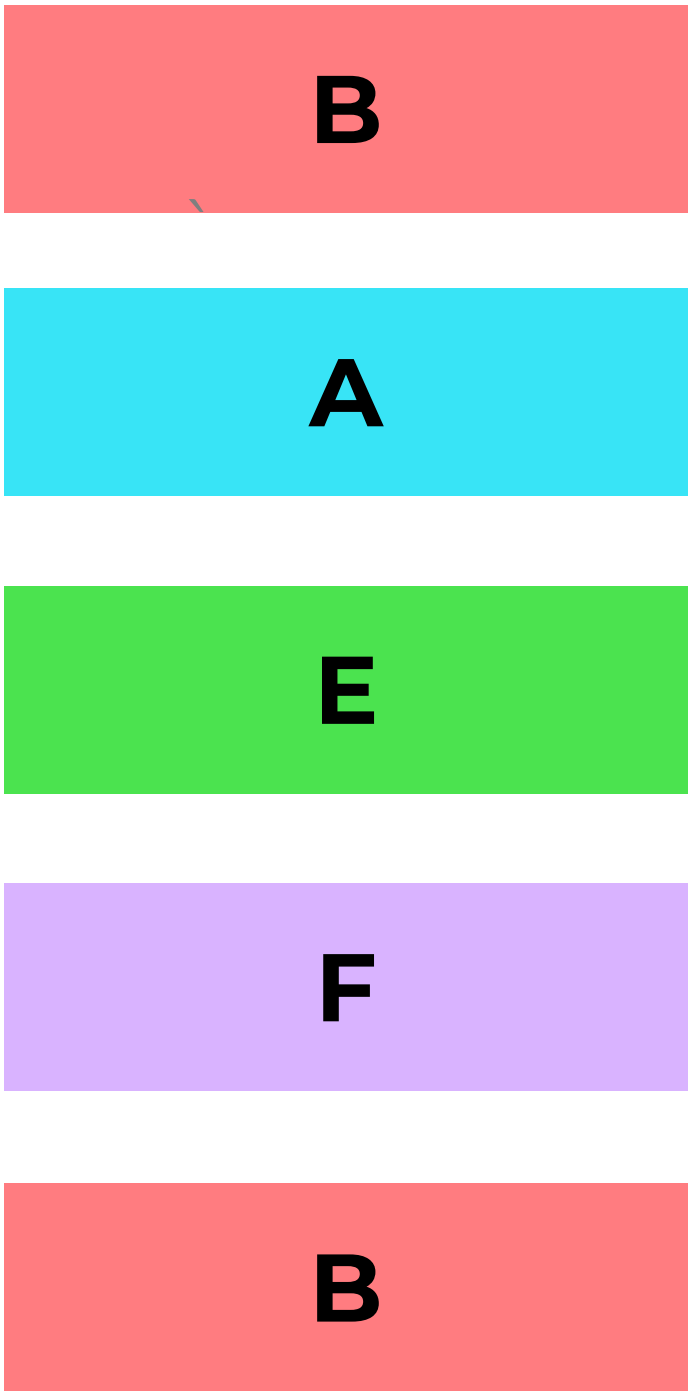
Let's write our own hash join!

HASH JOIN THEORY

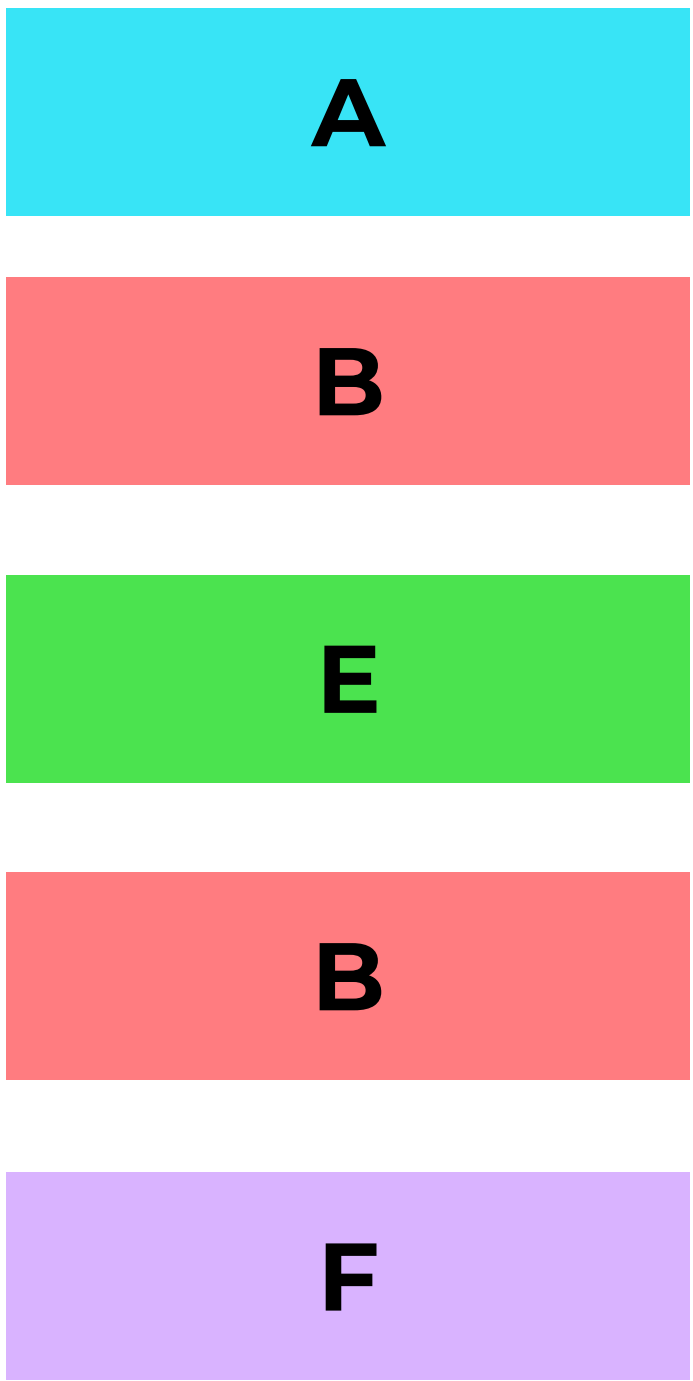
<https://bertwagner.com/posts/hash-match-join-internals/>



Build Input
(small dataset)



Probe Input
(large dataset)



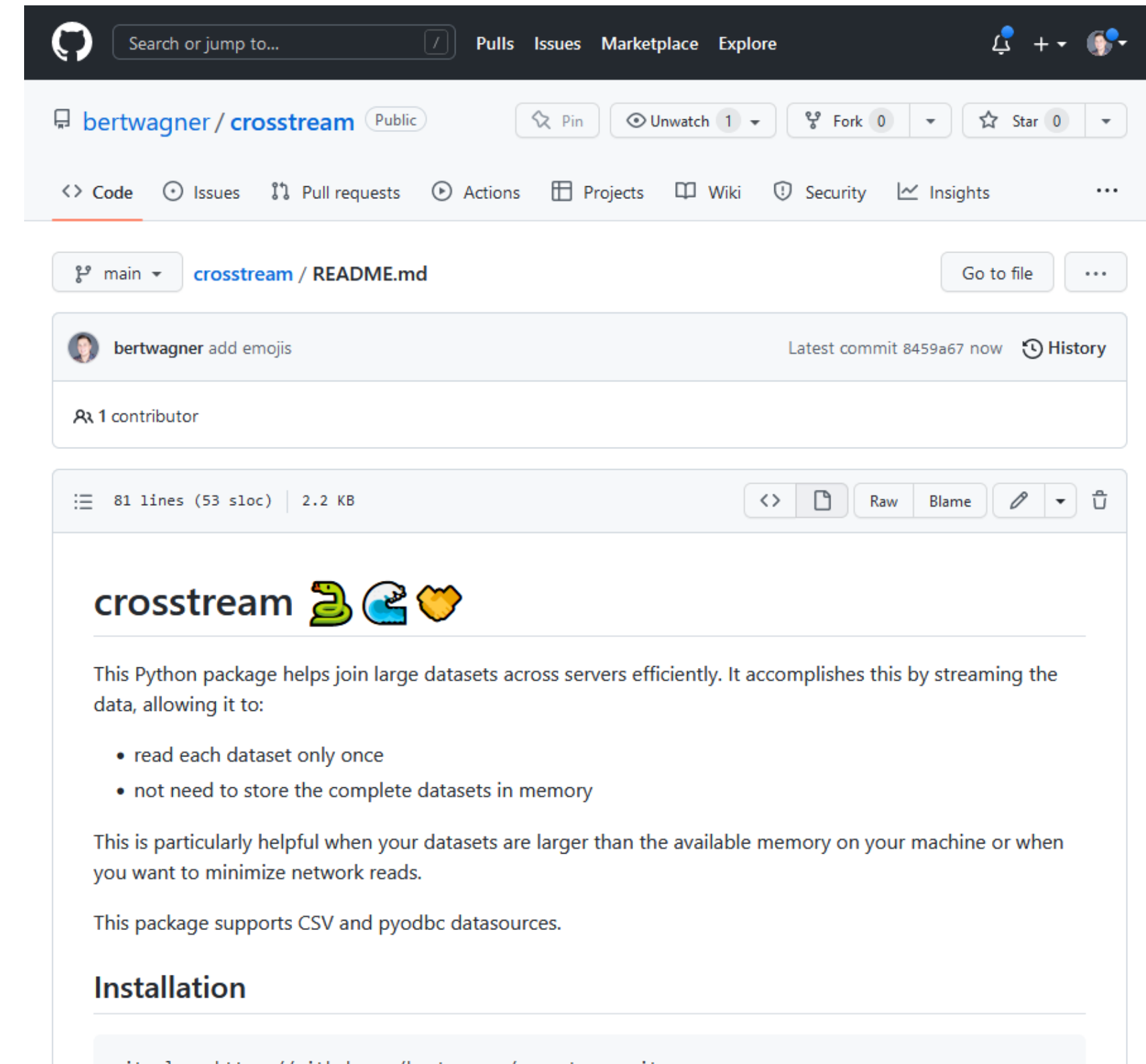
CROSSTREAM

Install from PyPI:

```
pip install crosstream
```

Or from source:

```
git clone https://github.com/bertwagner/crosstream.git
cd crosstream
pip install .
```



github.com/bertwagner/crosstream

DEMO:

BASIC USAGE

```
import crosstream as cs
import csv
```

```
file1 = 'small_dataset.csv'
file2 = 'large_dataset.csv'
```

```
# join using column indexes or column names
c1 = cs.read_csv(file1, True, [0, 1])
c2 = cs.read_csv(file2, True, ['col1', 'col2'])
```

```
# specify the output file
with open('joined_output.csv', 'w') as f:
    w = csv.writer(f)
```

```
# write header column names
w.writerow(c1.column_names + c2.column_names)
```

```
for row_left, row_right in cs.inner_hash_join(c1, c2):
    # write matched results to our joined_output.csv
    w.writerow(row_left + row_right)
```


DEMO:

CUSTOM
JOIN
KEYS

```
# define a function for transforming join key data before it's hashed
def custom_join_key_transform(value):
    transformed = value.replace(' ', '')
    return transformed

...
for row_left, row_right in cs.inner_hash_join(c1, c2,
                                              override_build_join_key=custom_join_key):
    # write matched results to our joined_output.csv
    w.writerow(row_left + row_right)
...
```


DEMO:

CUSTOM MATCH PROCESSING

```
# define a function for performing additional transformations or
# adding additional outputs before the columns are returned
def custom_process_matched_hashes(bucket_row,probe_row,
                                   bucket_join_column_indexes, probe_join_column_indexes):

    # adding a new column indicating the weights of these matches
    # are equal to 1
    weight=1.0
    return tuple(bucket_row),tuple(probe_row),(weight,)

...
for row_left,row_right in cs.inner_hash_join(c1,c2,
                                              override_process_matched_hashes=custom_process_matched_hashes):

    # write matched results to our joined_output.csv
    w.writerow(row_left + row_right)

...
```


SUMMARY: UGLY SOLUTION

- Last resort option
- Slow but reliable
- Can be programmed to restart on network failures
- Allows for heavy data transformations before and after joining
- Reads each dataset only once
- Works on CSV and ODBC
- Assuming you can't fit data in memory or on your laptop's disk
 - If only memory constrained, can use [Dask to swap dataframes from disk](#)

THANK YOU!



@bertwagner



bertwagner.com



youtube.com/DataWithBert



bertwagner@bertwagner.com