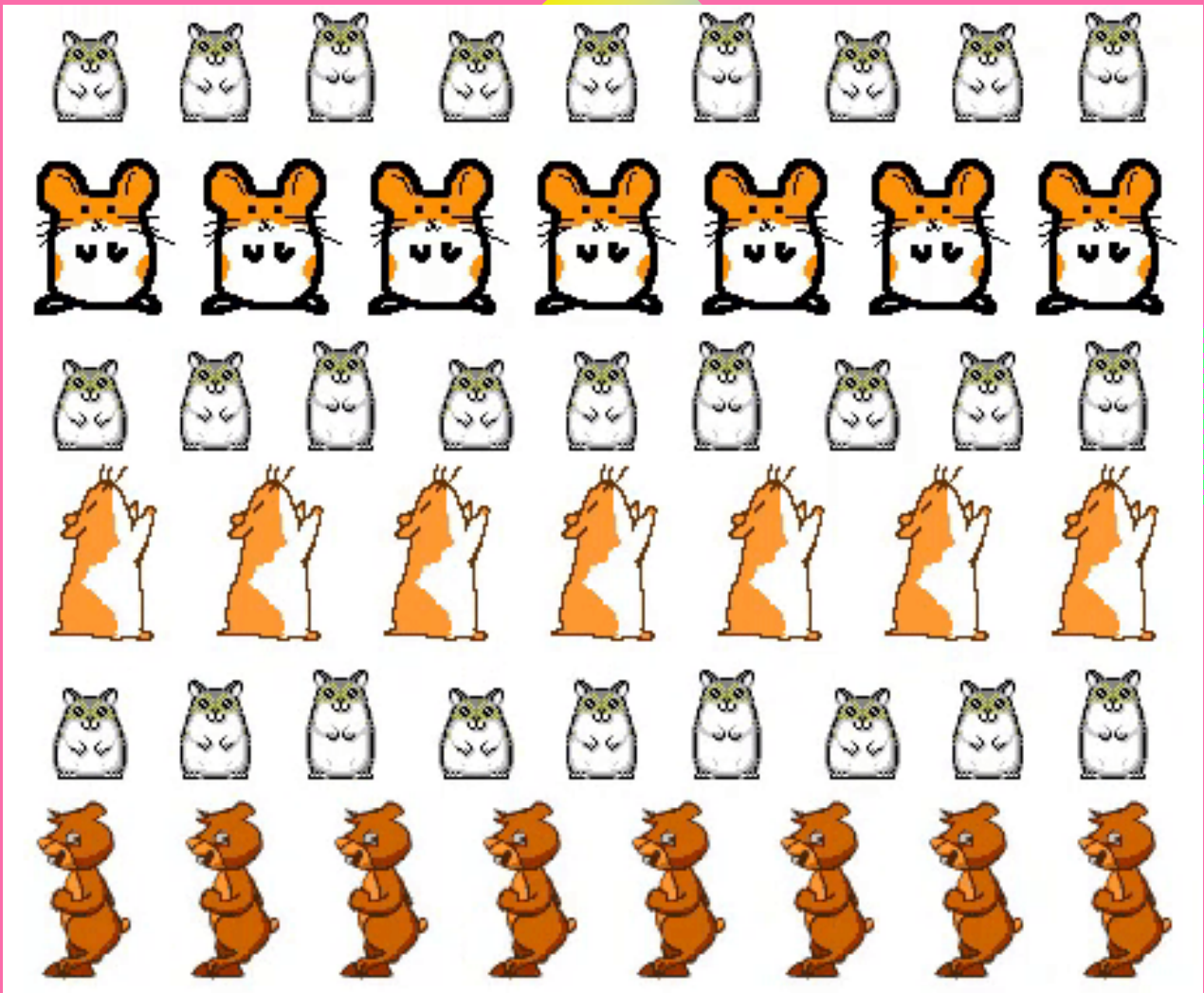
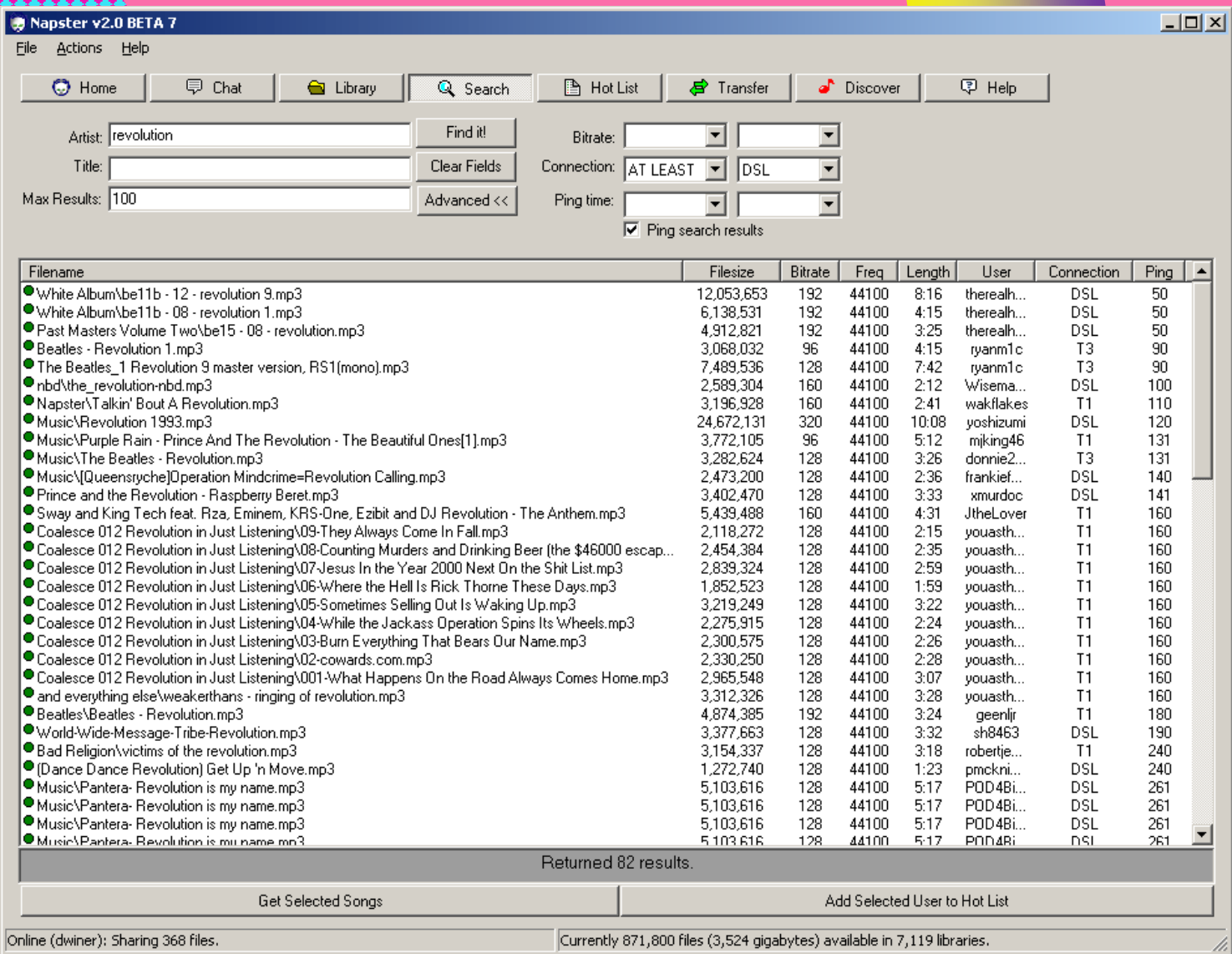



Cross-Server Data Joins on Slow Networks with Python

Bert Wagner
PyCon 2023


“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.

<div style="display: flex; justify-content: space-between;"> nd February 09, 2005: SUBMIT A LINK # of Comments </div>
<div style="display: flex; justify-content: space-between; align-items: flex-start;"> <div style="width: 10%;"> (Some Guy) </div> <div style="width: 10%; text-align: center;"> <div style="background-color: #f08080; padding: 2px; font-size: 8px;">INTERESTING</div> </div> <div style="width: 70%;"> Instead of making taxpayers foot the bill for prison costs, why not let the prisoners pay for it themselves? </div> <div style="width: 10%; text-align: right;"> (64) </div> </div>
<div style="display: flex; justify-content: space-between; align-items: flex-start;"> <div style="width: 10%;"> (Some Demolition Site) </div> <div style="width: 10%; text-align: center;"> <div style="background-color: #4682b4; color: white; padding: 2px; font-size: 8px;">Photoshop</div> </div> <div style="width: 70%;"> Photoshop this building, or how it got this way </div> <div style="width: 10%; text-align: right;"> (75) </div> </div>
<div style="display: flex; justify-content: space-between; align-items: flex-start;"> <div style="width: 10%;">  </div> <div style="width: 10%; text-align: center;"> <div style="background-color: #32cd32; padding: 2px; font-size: 8px;">WORLD</div> </div> <div style="width: 70%;"> 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 style="width: 10%; text-align: right;"> (15) </div> </div>
<div style="display: flex; justify-content: space-between; align-items: flex-start;"> <div style="width: 10%;">  </div> <div style="width: 10%; text-align: center;"> <div style="background-color: #32cd32; padding: 2px; font-size: 8px;">WORLD</div> </div> <div style="width: 70%;"> A record 2,201 cases of "train molestation" were reported in Tokyo last year </div> <div style="width: 10%; text-align: right;"> (32) </div> </div>
<div style="display: flex; justify-content: space-between; align-items: flex-start;"> <div style="width: 10%;"> (SiLive.com) </div> <div style="width: 10%; text-align: center;"> <div style="background-color: #8b4513; color: white; padding: 2px; font-size: 8px;">OBVIOUS</div> </div> <div style="width: 70%;"> Homeless man freezes to death after seeking shelter in abandoned ice factory </div> <div style="width: 10%; text-align: right;"> (24) </div> </div>
<div style="display: flex; justify-content: space-between; align-items: flex-start;"> <div style="width: 10%;">  </div> <div style="width: 10%; text-align: center;"> <div style="background-color: #ff4500; color: white; padding: 2px; font-size: 8px;">AMUSING</div> </div> <div style="width: 70%;"> The 10 most disrespected entities in sports today </div> <div style="width: 10%; text-align: right;"> (40) </div> </div>
<div style="display: flex; justify-content: space-between; align-items: flex-start;"> <div style="width: 10%;">  </div> <div style="width: 10%; text-align: center;"> <div style="background-color: #4682b4; color: white; padding: 2px; font-size: 8px;">UNLIKELY</div> </div> <div style="width: 70%;"> 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 style="width: 10%; text-align: right;"> (29) </div> </div>
<div style="display: flex; justify-content: space-between; align-items: flex-start;"> <div style="width: 10%;"> (nbc4.tv) </div> <div style="width: 10%; text-align: center;"> <div style="background-color: #8b4513; color: white; padding: 2px; font-size: 8px;">STRANGE</div> </div> <div style="width: 70%;"> This week's unknown nail in the body belongs to a man with a pain in the neck </div> <div style="width: 10%; text-align: right;"> (31) </div> </div>
<div style="display: flex; justify-content: space-between; align-items: flex-start;"> <div style="width: 10%;">  </div> <div style="width: 10%; text-align: center;"> <div style="background-color: #483d8b; color: white; padding: 2px; font-size: 8px;">SAD</div> </div> <div style="width: 70%;"> Doobie Brothers drummer Keith Knudsen dies at 56 </div> <div style="width: 10%; text-align: right;"> (48) </div> </div>
<div style="display: flex; justify-content: space-between; align-items: flex-start;"> <div style="width: 10%;"> (Gallup) </div> <div style="width: 10%; text-align: center;"> <div style="background-color: #f08080; padding: 2px; font-size: 8px;">INTERESTING</div> </div> <div style="width: 70%;"> Bush's approval rating hits 57 percent. Farkers' heads asplode </div> <div style="width: 10%; text-align: right;"> (324) </div> </div>

[HELP](#)
[FORUMS](#)
[CHAT](#)
[LITERATURE](#)
[WEBCAMS](#)
[USERS](#)
[NETWORK](#)
[SUBMIT CONTENT](#)

[HTTP://WWW.NEUGROUND](#)

LATEST 5 SUBMISSIONS

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

NG FEATURES!

1. Dating Sims
2. Dress-Up Games
3. Britney Spears
4. Webcams
5. Car & Driving Games
6. Celebrity Killing

FEATURED NG MEDIA!

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!

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

[FRONT PAGE ARCHIVE >>](#)
[1,000'S MORE IN THE PORTAL!](#)

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!

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!

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




digg

login

latest front page stories

207
diggs

[PSP vs DS one year later.](#)


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

[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

["Easy Button" Hack](#)


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

[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

[The Top 10 weirdest keyboards ever](#)

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

[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.](#)

[25 comments](#) | [blog this](#) | [email this](#) | category: [gadgets](#)



Broadcast Yourself™


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

Videos


Categories

Channels


Community


[Upload Videos](#)

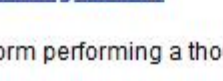
Director Videos




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



[Hometown Baghdad...](#)
chattheplanet



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



[Black Anthem](#)
pdrog

[See More Featured Videos](#)

Member Login


YouTube Username:

YouTube Password:

[Sign Up](#)


[Forgot Username](#) | [Forgot Password](#)

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'

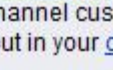
From: [whitestkidsdotcom](#) Views: 12,199 ★★★★★ More in [Comedy](#)

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.

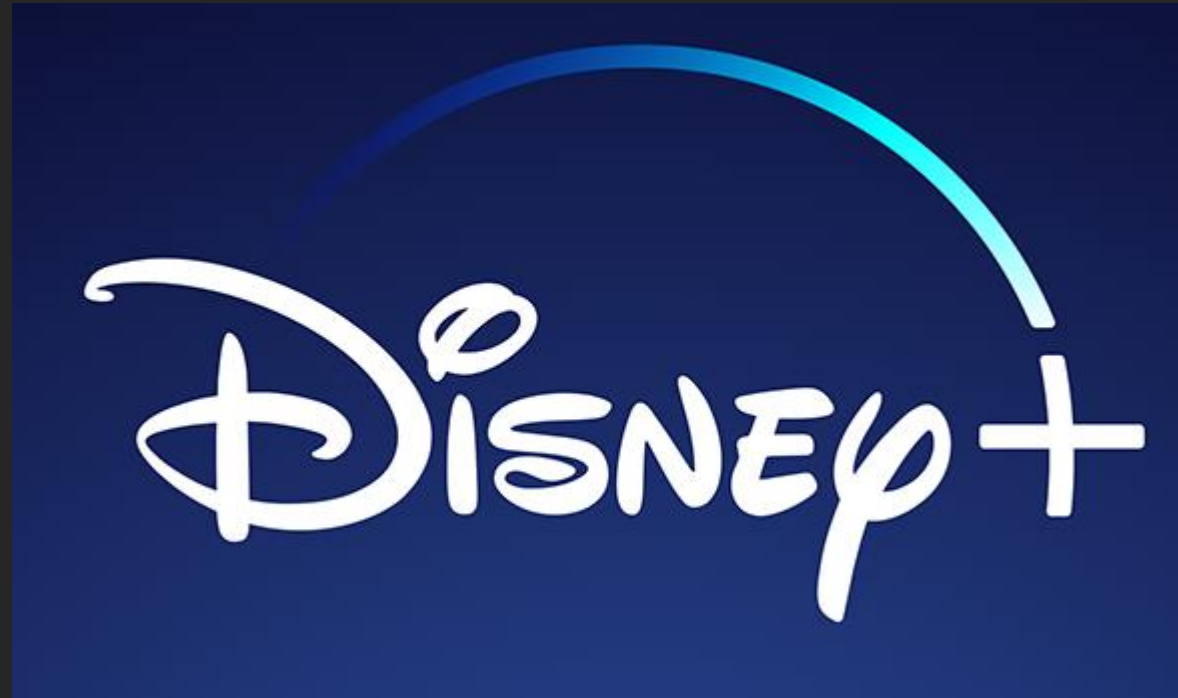


[TestTube](#)

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



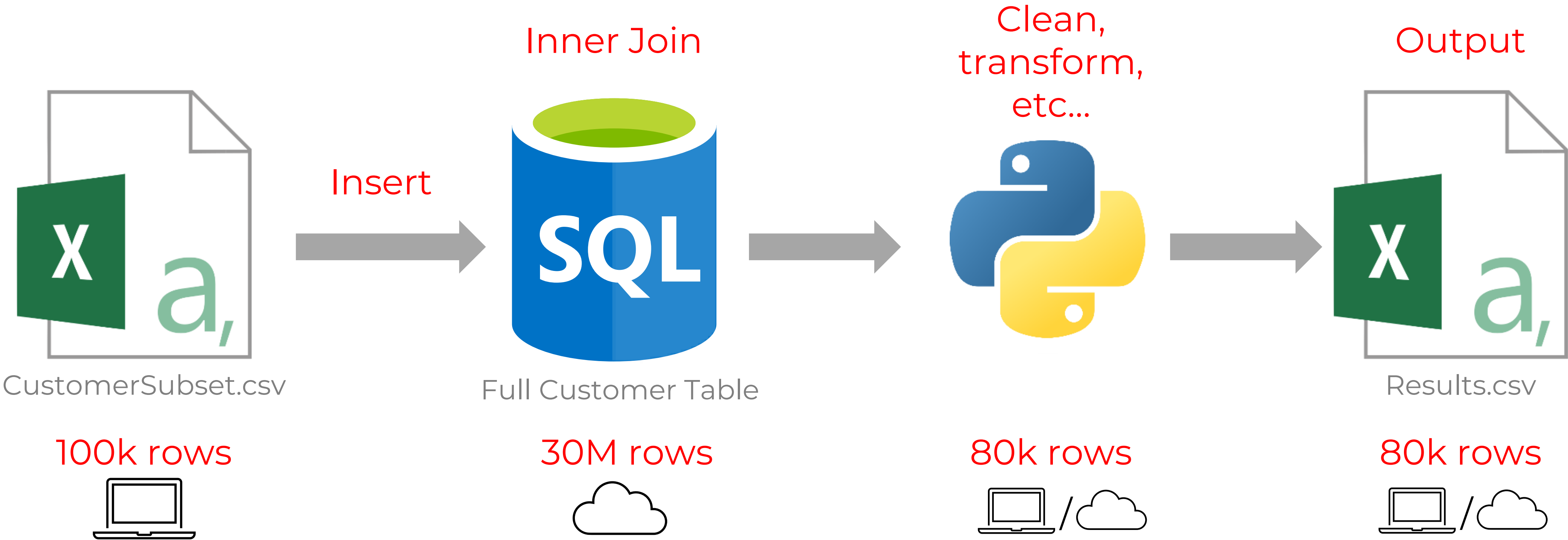
CustomerSubset.csv
100k rows
2GB
Local



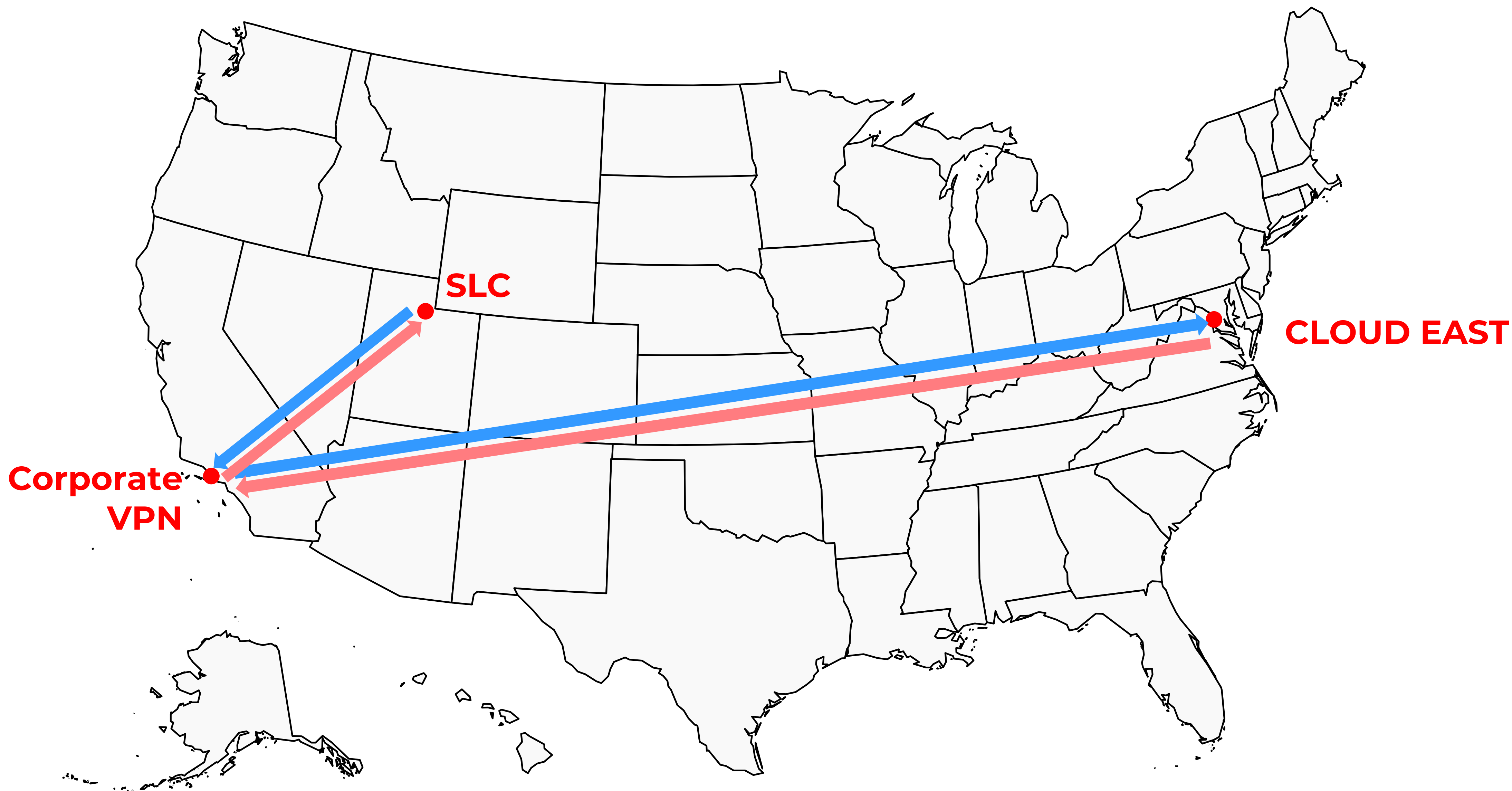
All Customer Data
30M rows
20GB
Remote

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](#)

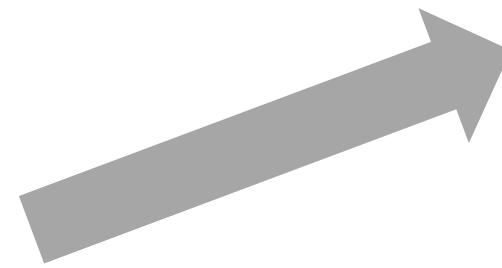
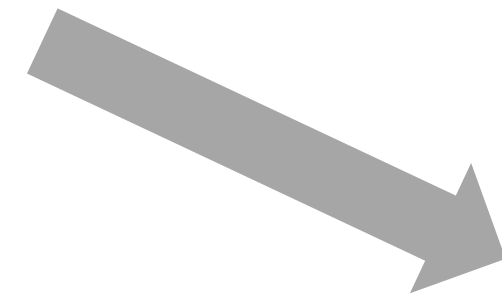
SOLUTION: NOT BAD



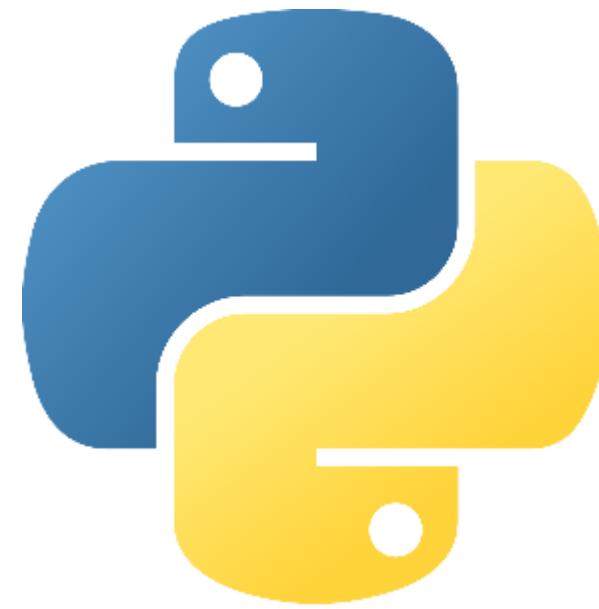
Full Customer Table



Customer
Subset.csv



Big heavy
duty server



Results.csv



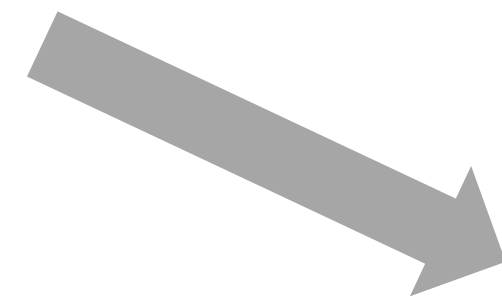
SOLUTION: NOT BAD



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:

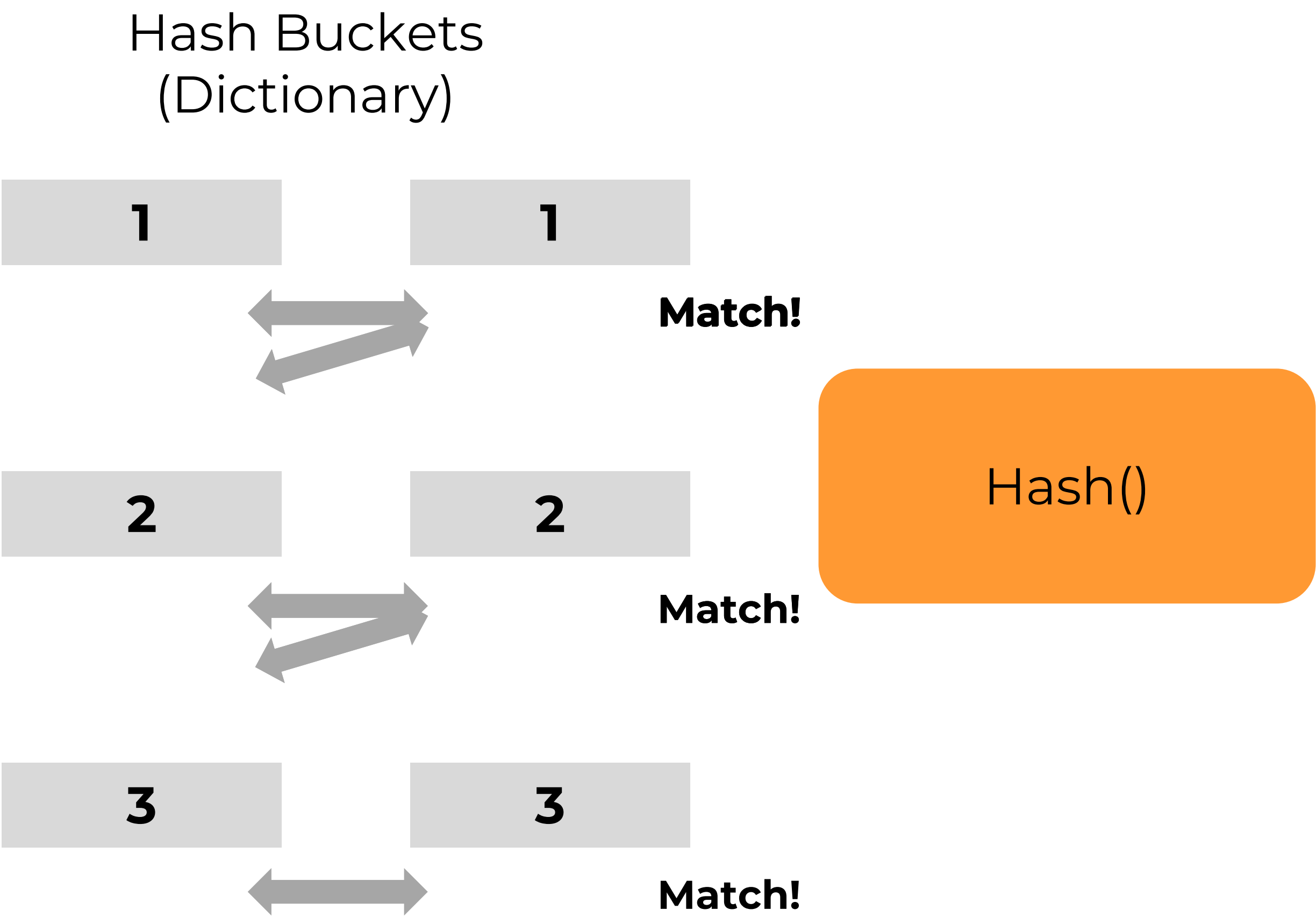
- [Pandas: Scaling to large datasets](#)

SOLUTION: UGLY

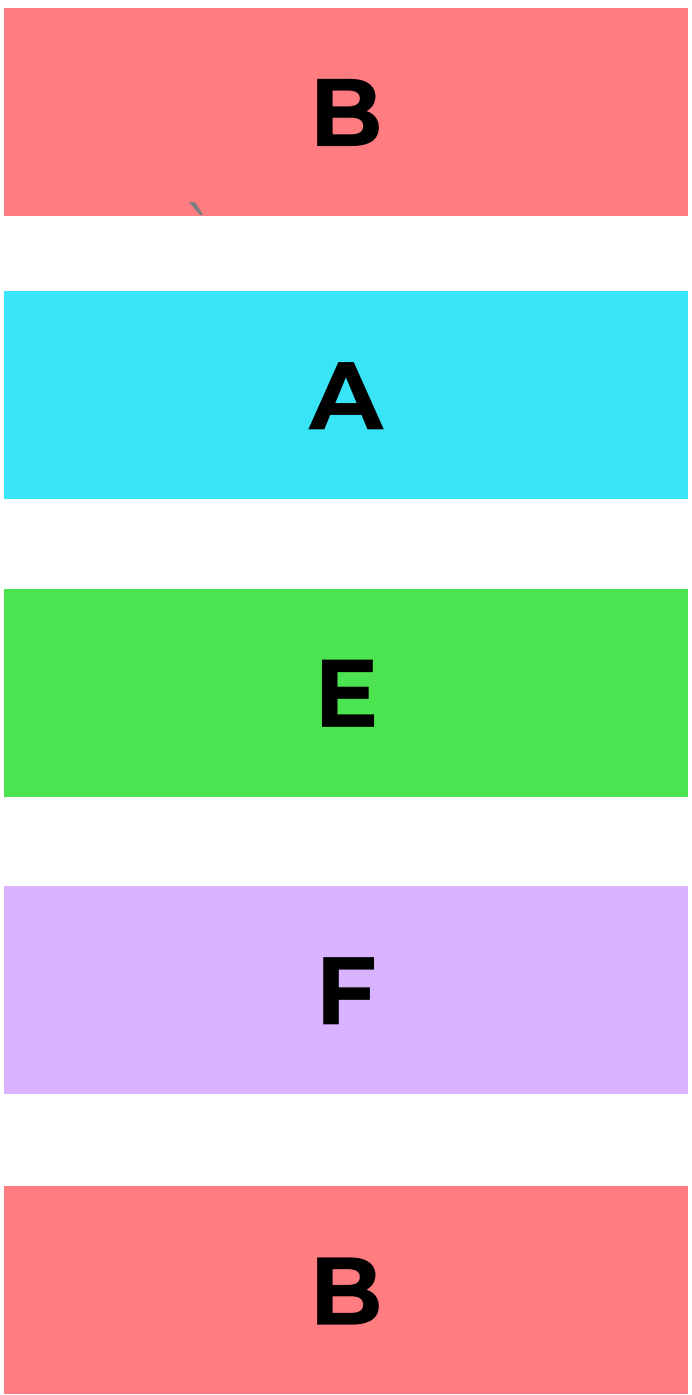
Let's write our own hash join!

HASH JOIN THEORY

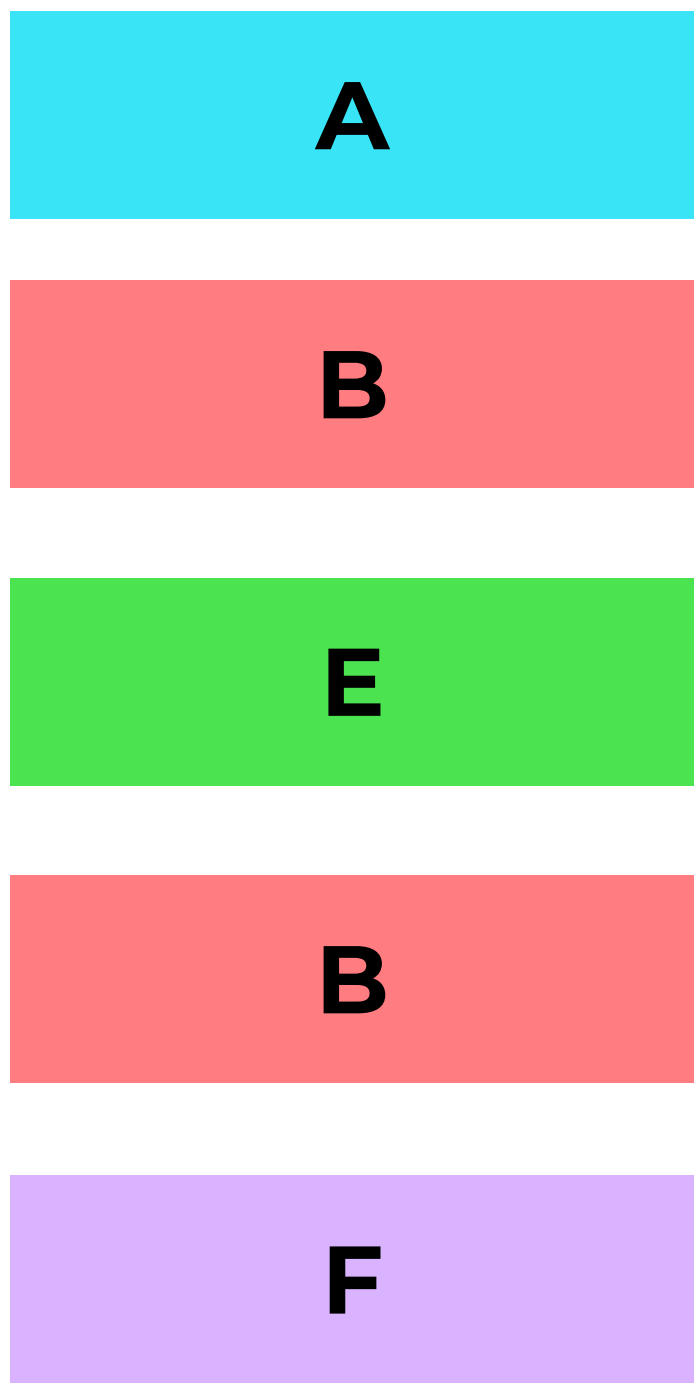
<https://bertwagner.com/posts/hash-match-join-internals/>



Build Input
(small dataset)



Probe Input
(large dataset)



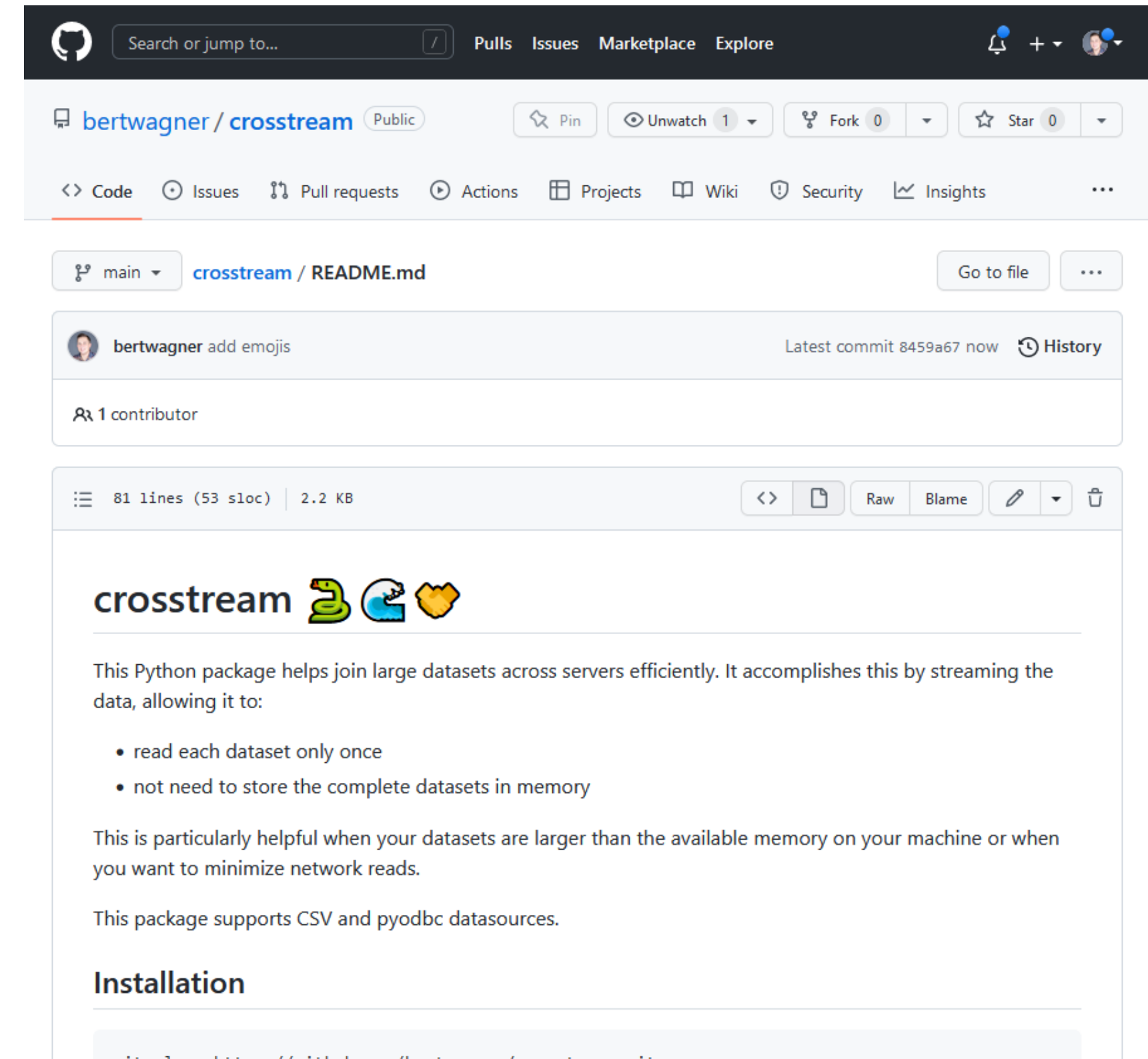
CROSSTREAM

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 joining on criteria that is modified before
# inserting into the hash table
def custom_join_key(row,indices):
    # calculate the hash of join values
    join_values = []
    join_key_values = []
    for col_index in indices:
        # here we transform our join key, removing any spaces from
        # our values
        col_value = str(row[col_index]).replace(' ','')
        join_values.append(str(hash(col_value)))
        join_key_values.append(col_value)
    join_key = ''.join(join_values)

    return join_key, join_key_values

...
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