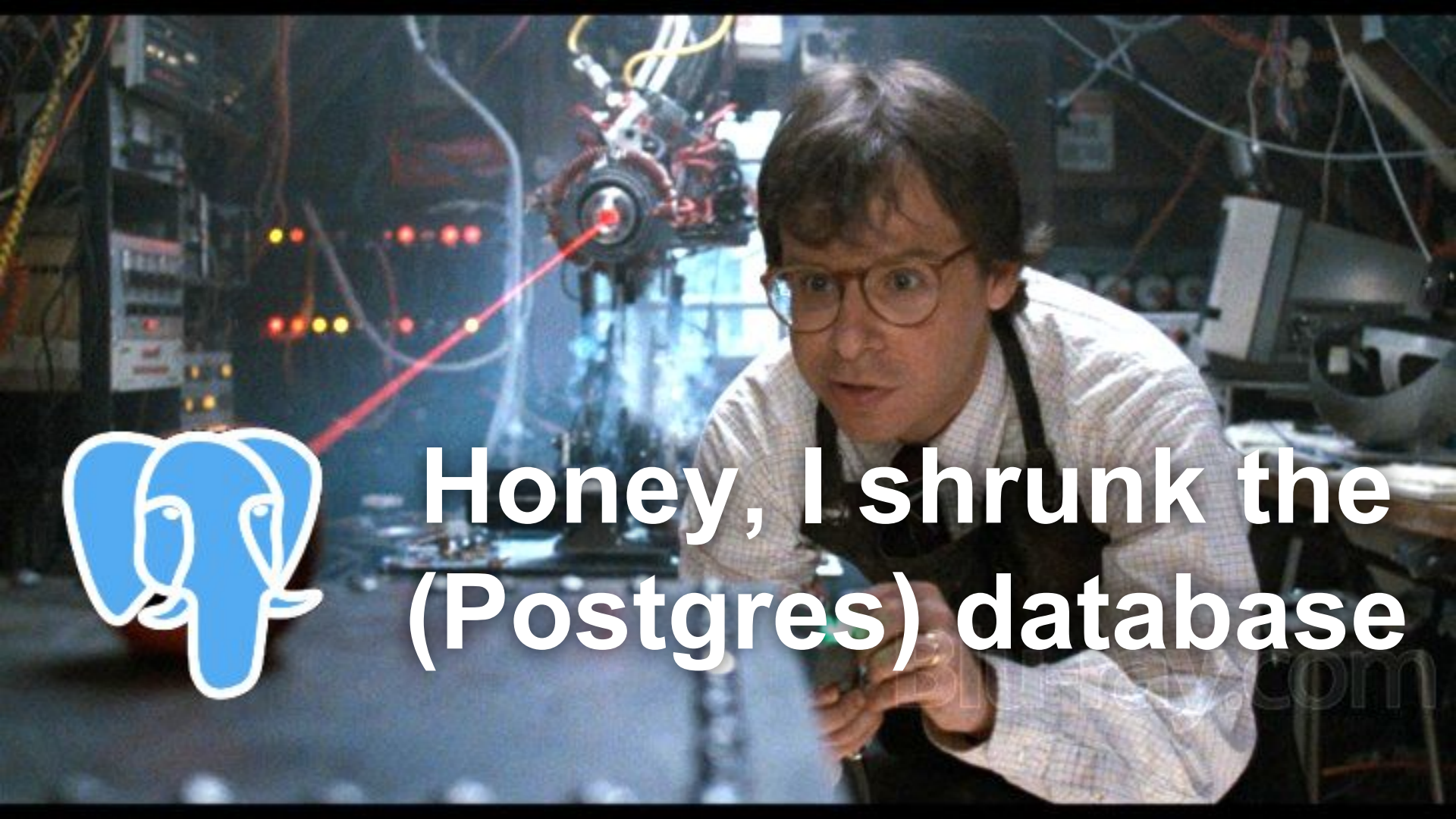
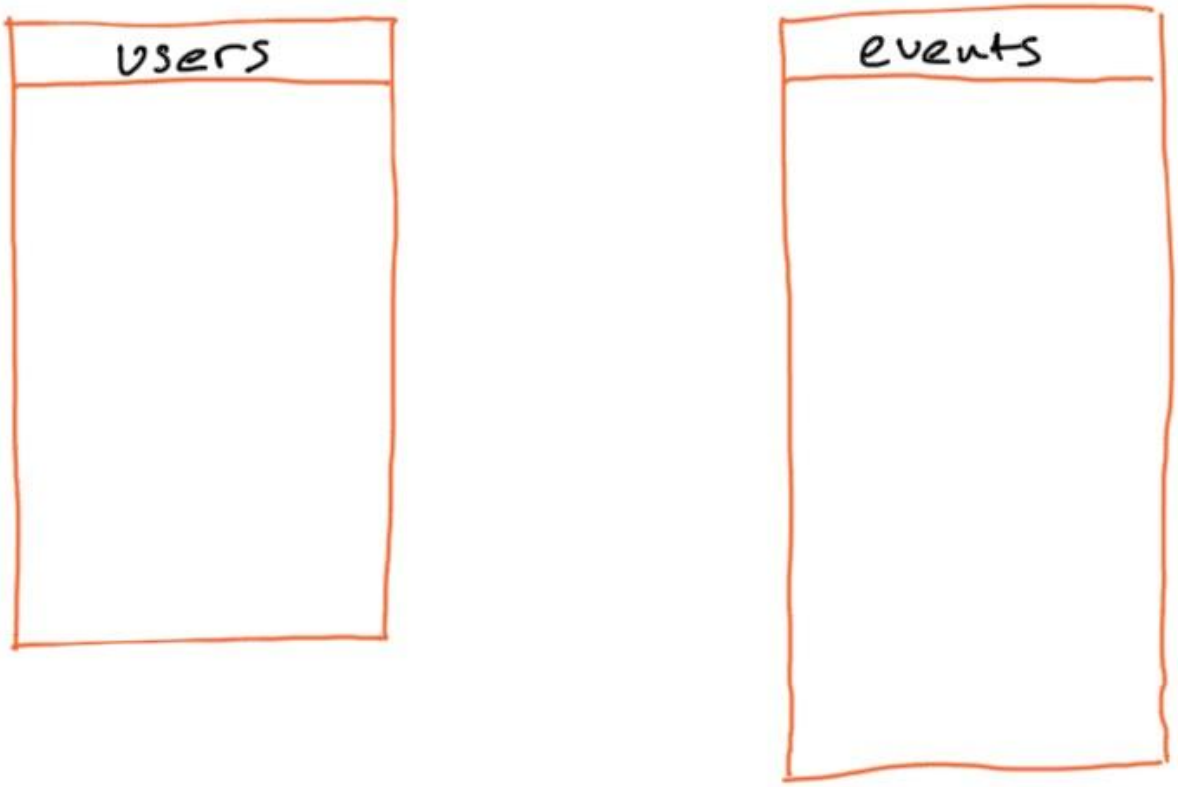

Honey, I shrunk the
(Postgres) database

What is database subsetting?

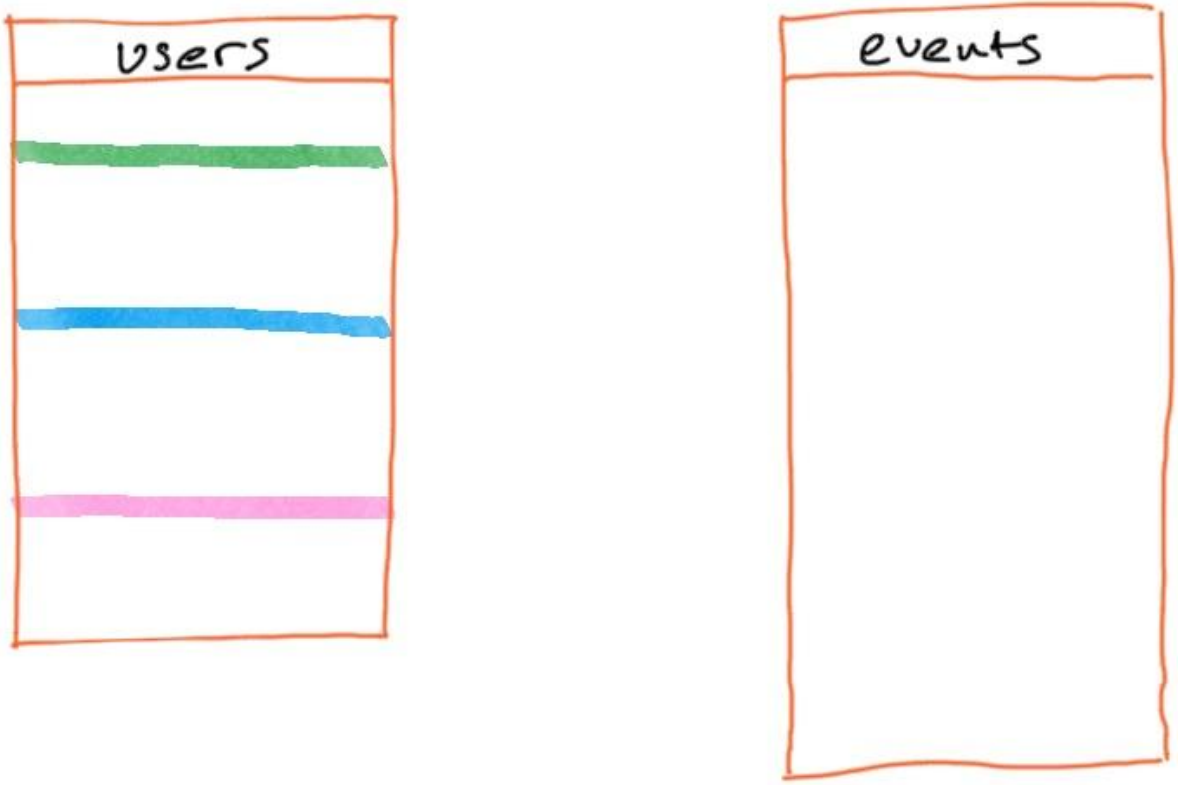
Taking a representative sample of a dataset in a manner that preserves the referential integrity of your database.

Example: give me 5% of my users

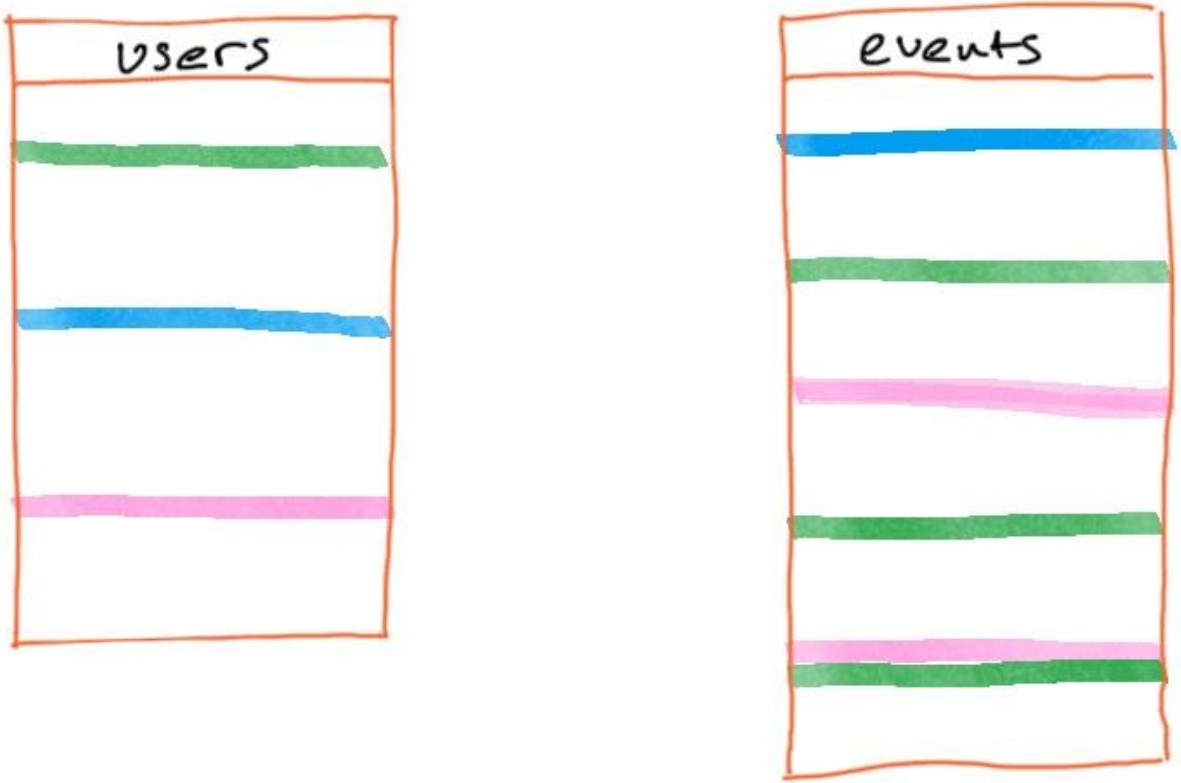
What is database subsetting?



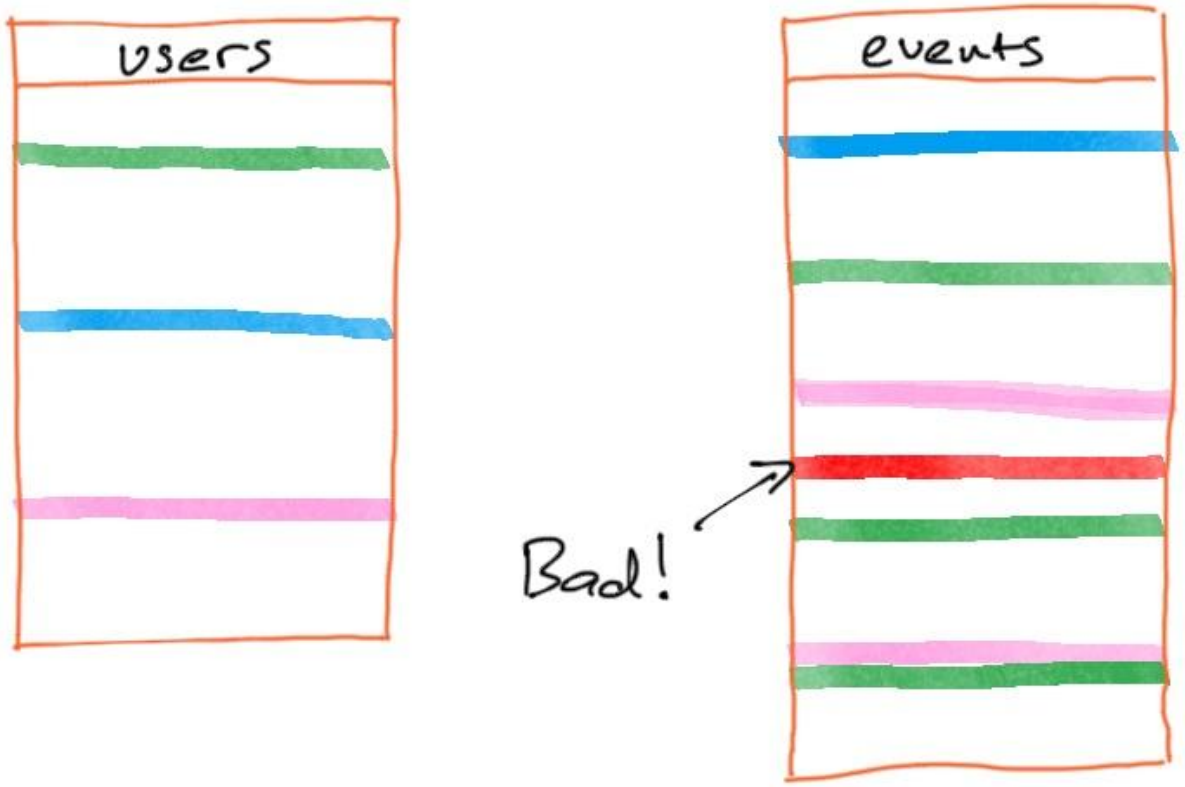
What is database subsetting?



What is database subsetting?



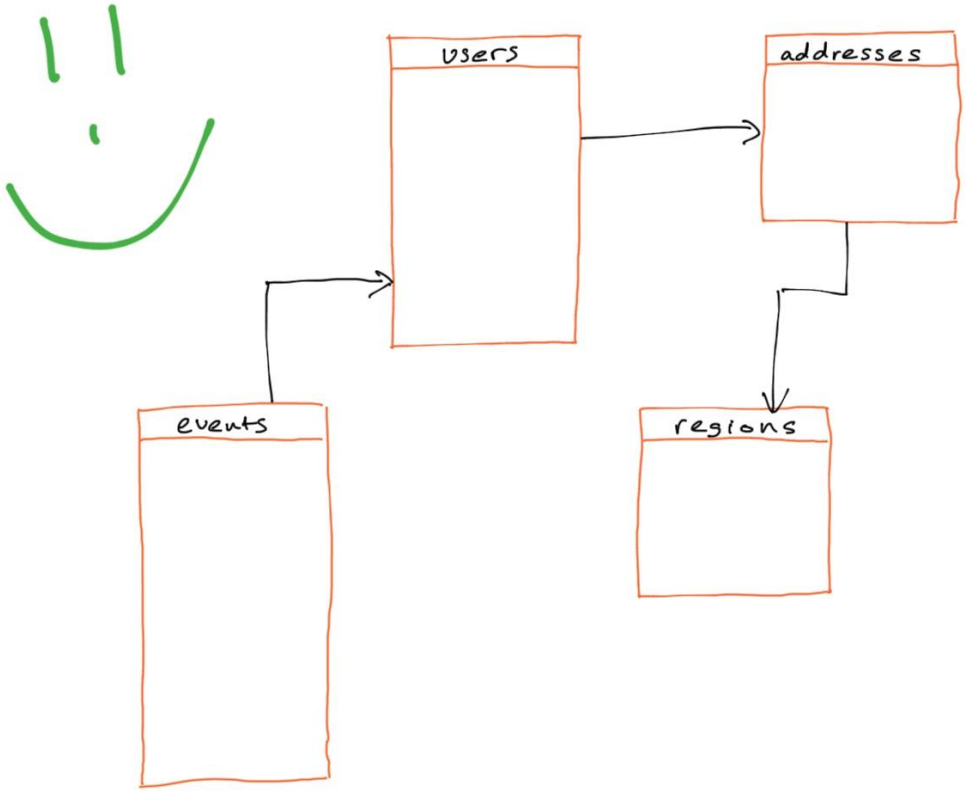
What is database subsetting?



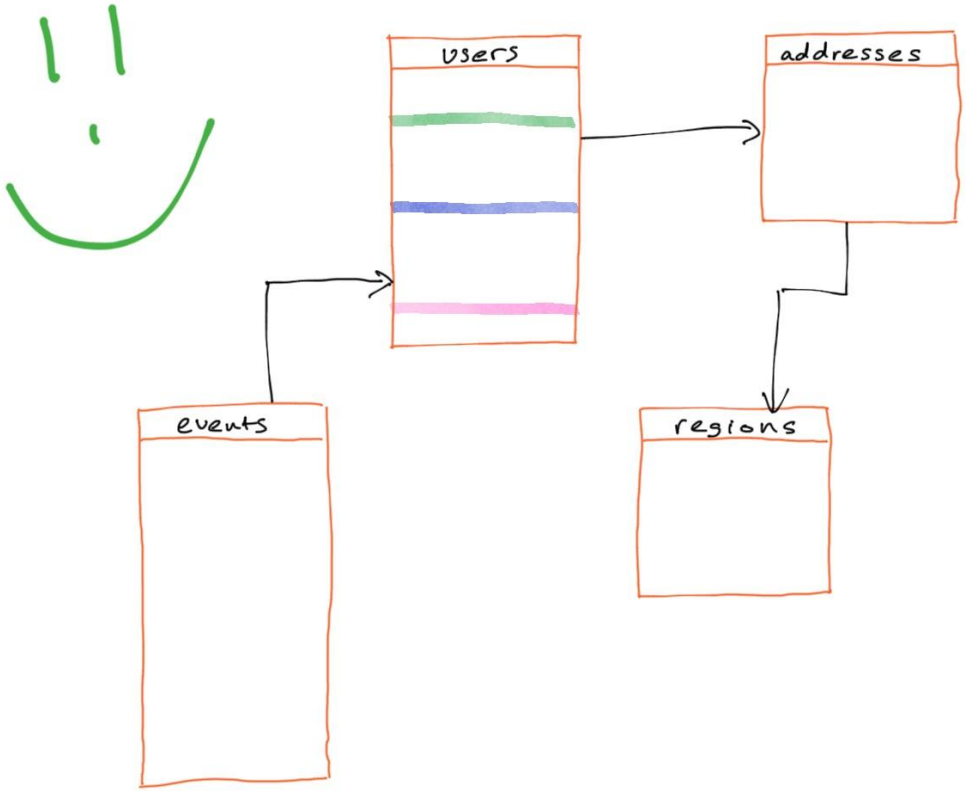
Why would someone want to subset a database?

1. Shrink a production database for cost effective staging/test.
2. Focus on a few specific rows (and related rows) to reproduce a bug.
3. Share a database with others when they’re not allowed to see parts of it.

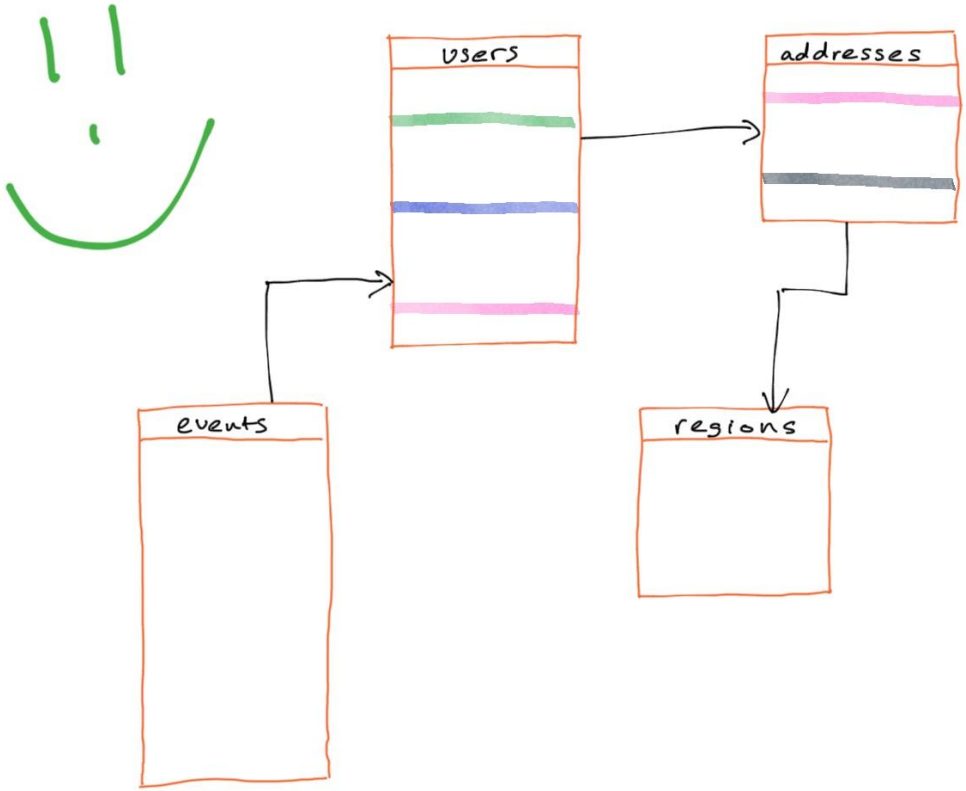
Approach 1: Random Subsetting



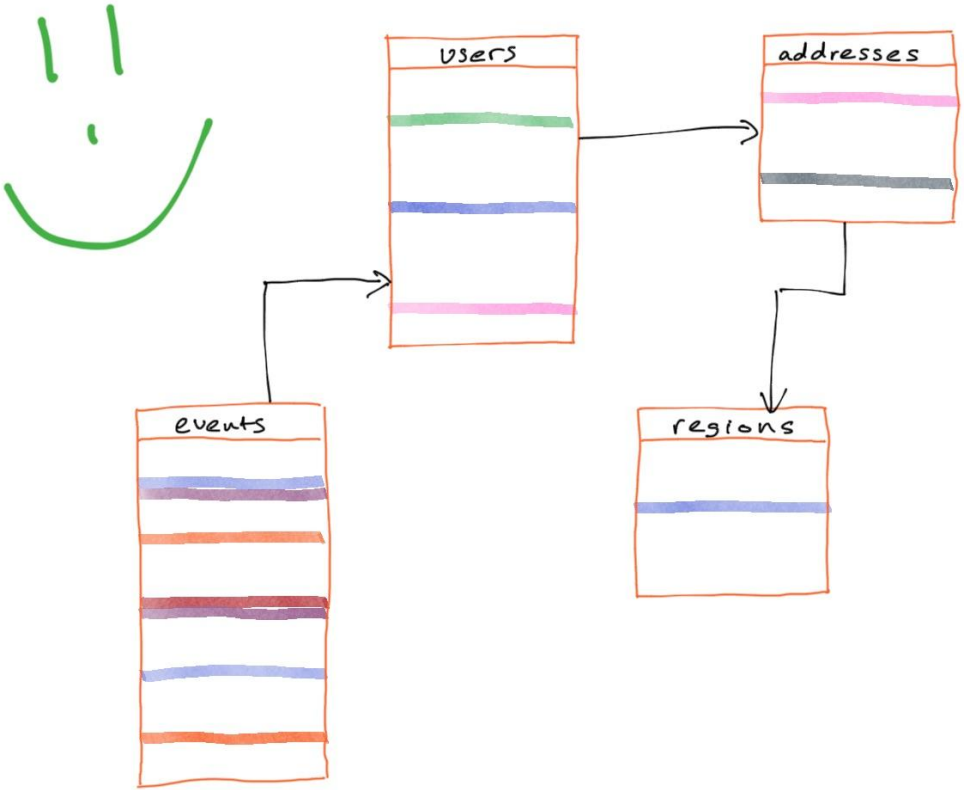
Approach 1: Random Subsetting



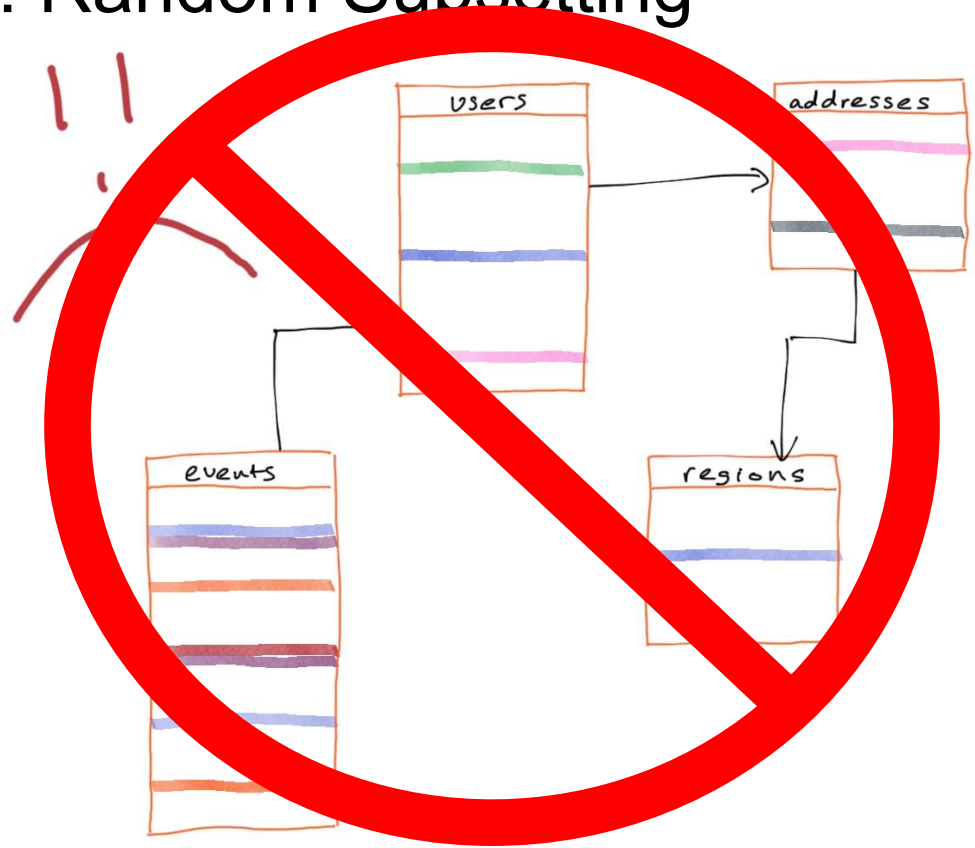
Approach 1: Random Subsetting



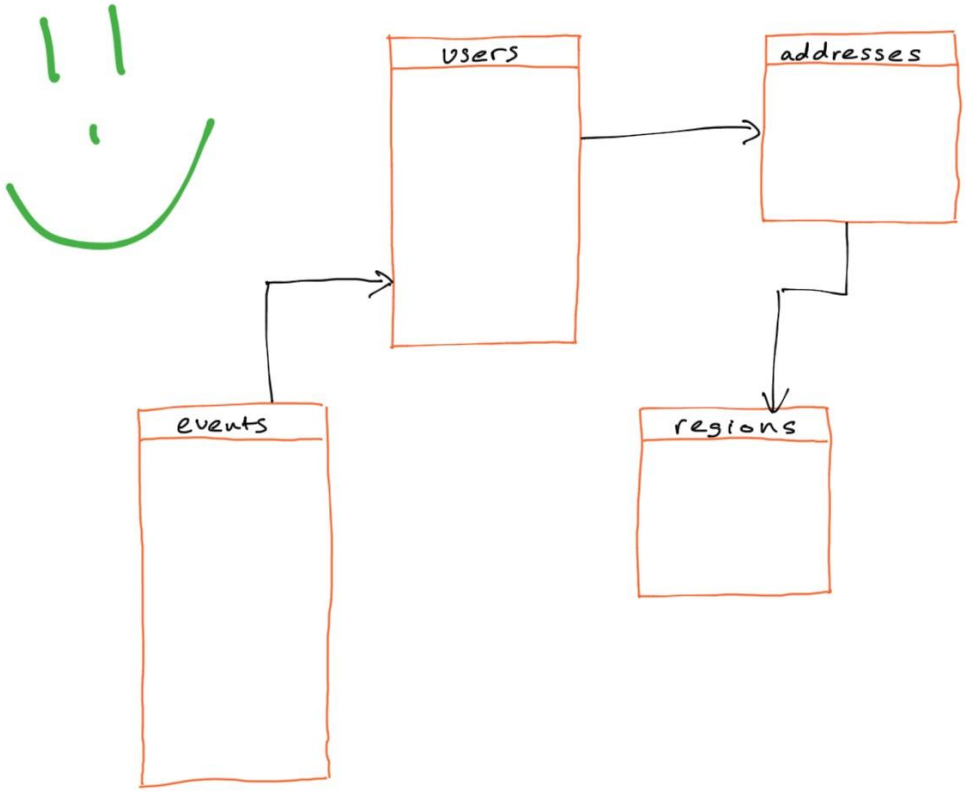
Approach 1: Random Subsetting



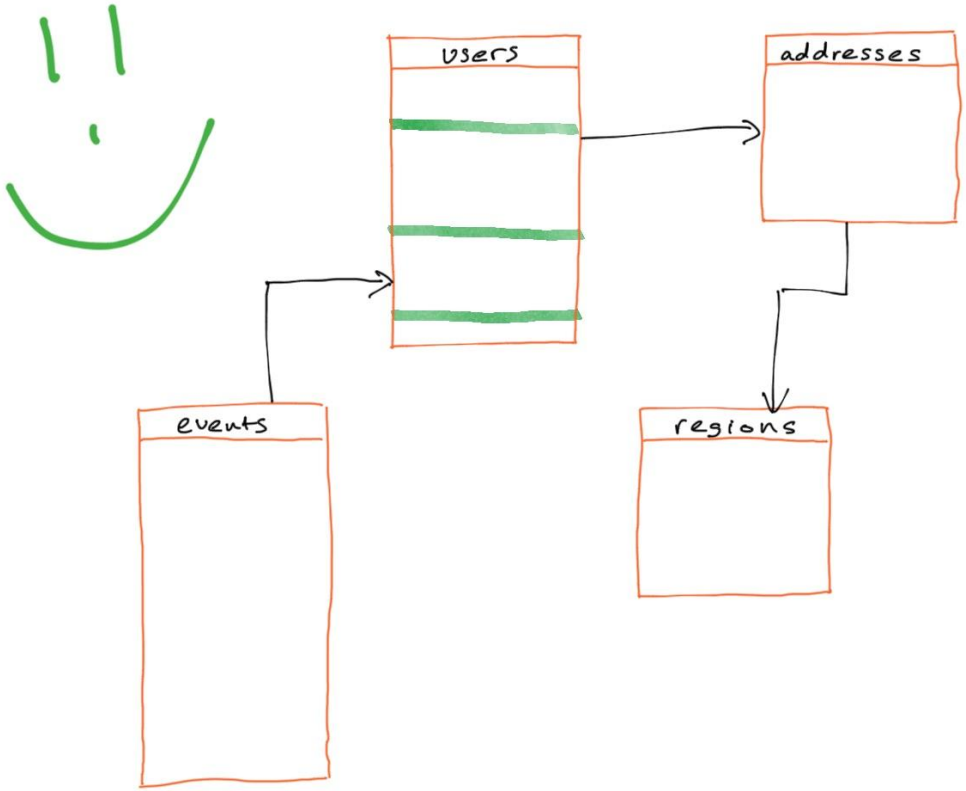
Approach 1: Random Subsetting



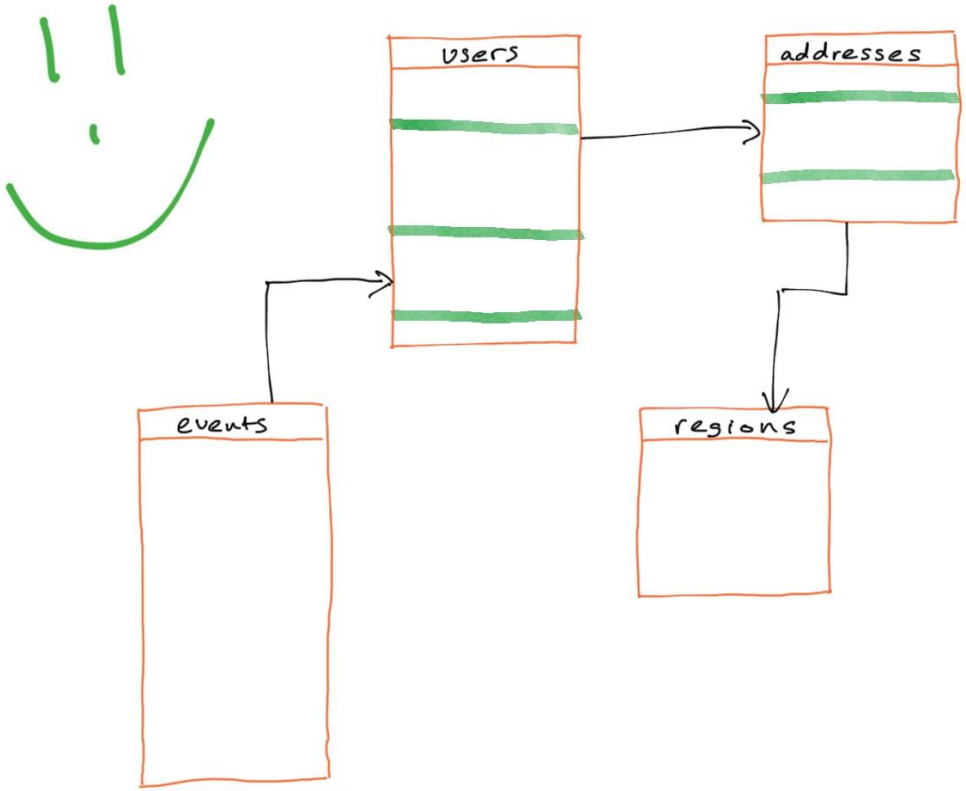
Approach 2: Follow the Foreign Keys



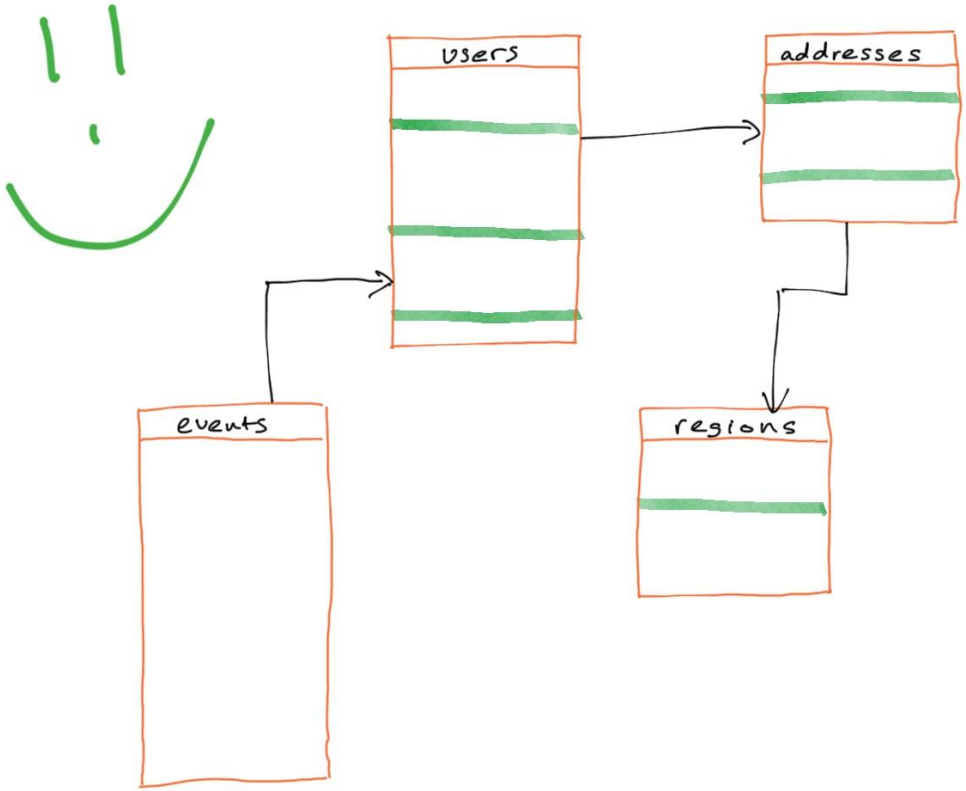
Approach 2: Follow the Foreign Keys



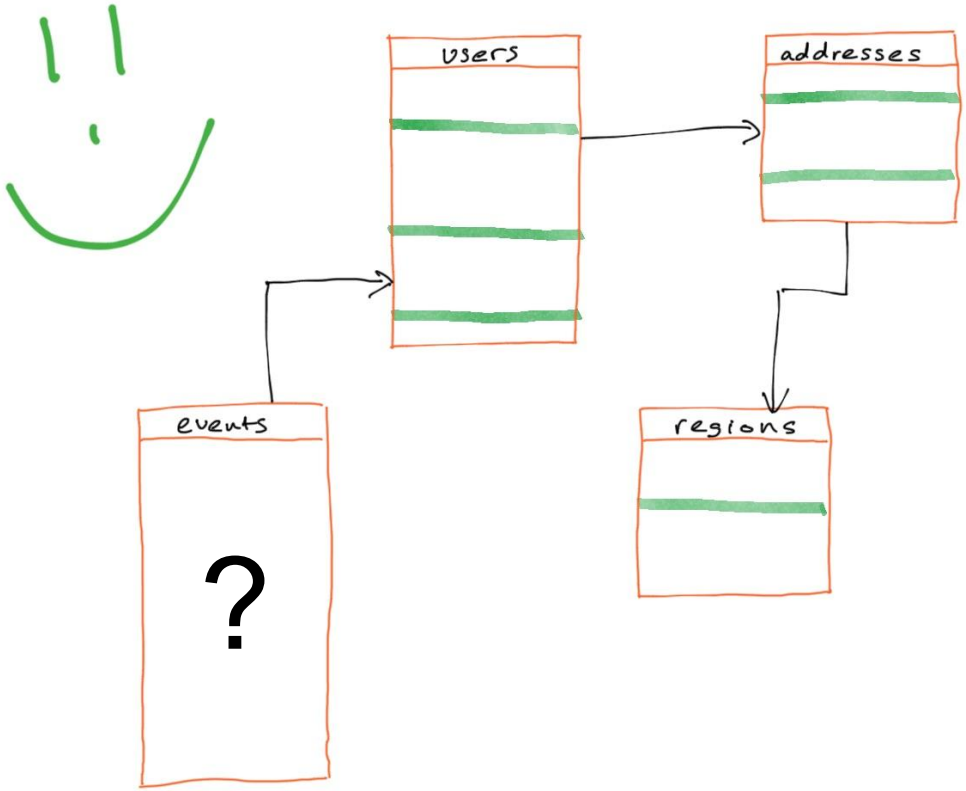
Approach 2: Follow the Foreign Keys



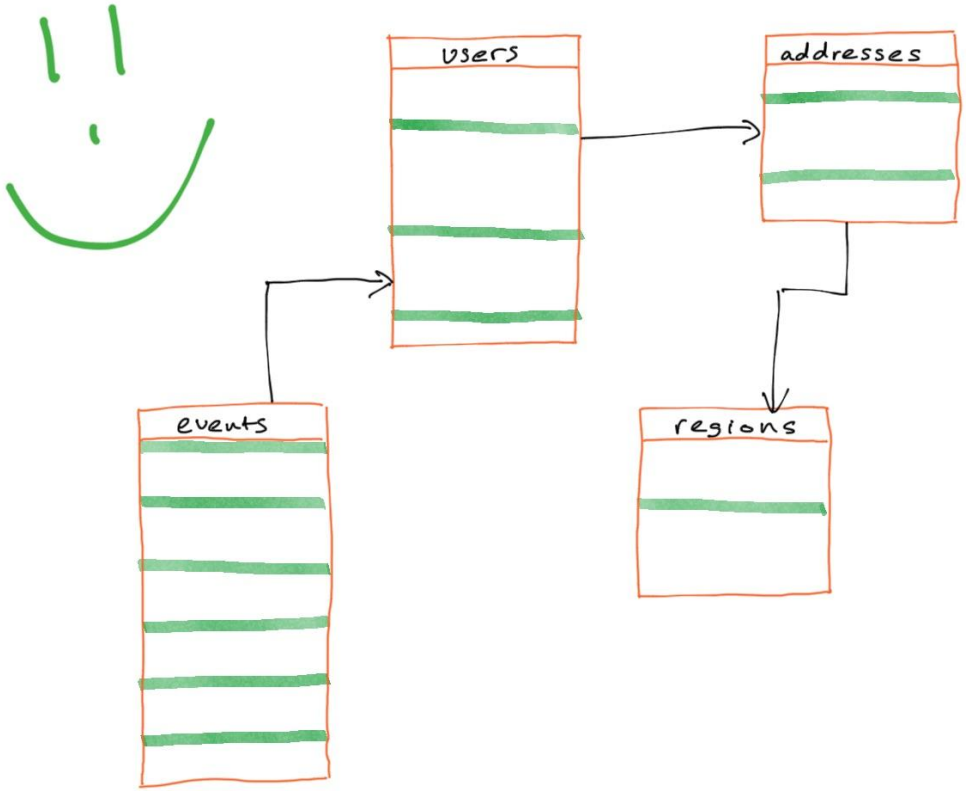
Approach 2: Follow the Foreign Keys



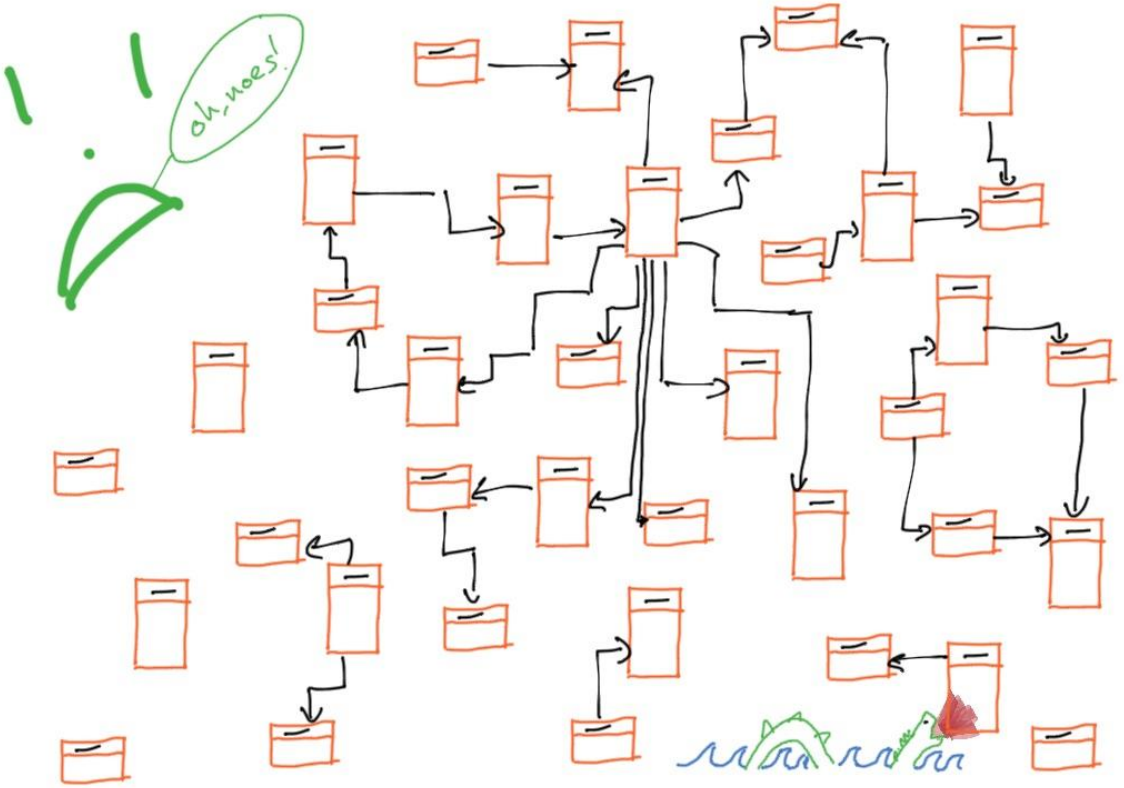
Approach 2: Follow the Foreign Keys



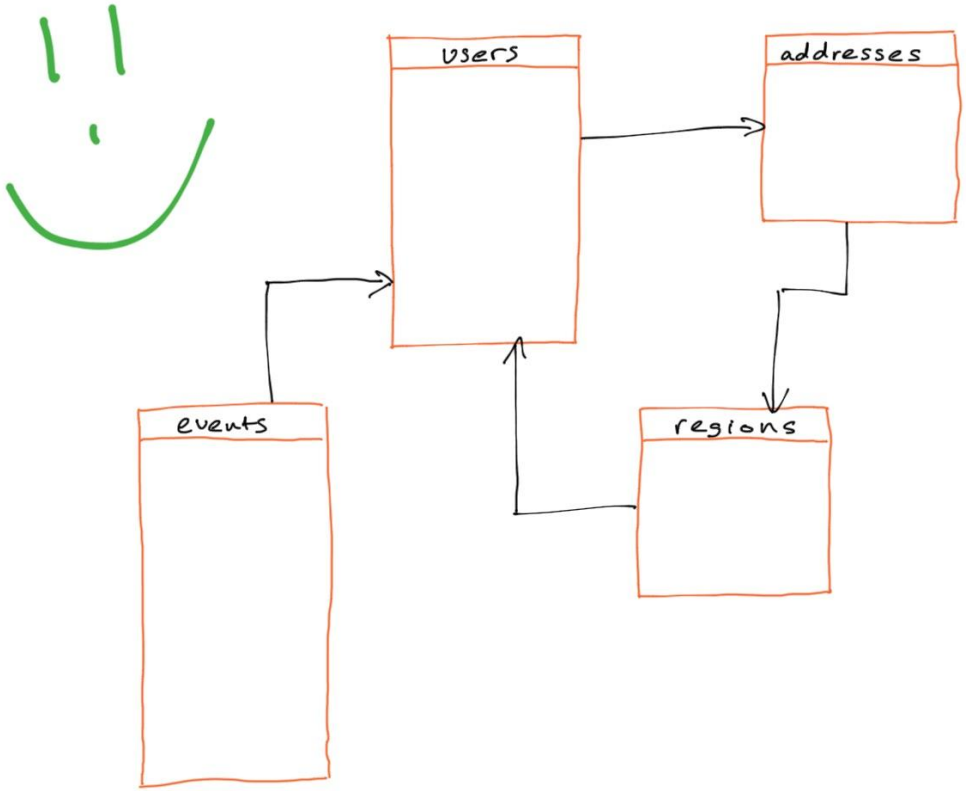
Approach 2: Follow the Foreign Keys



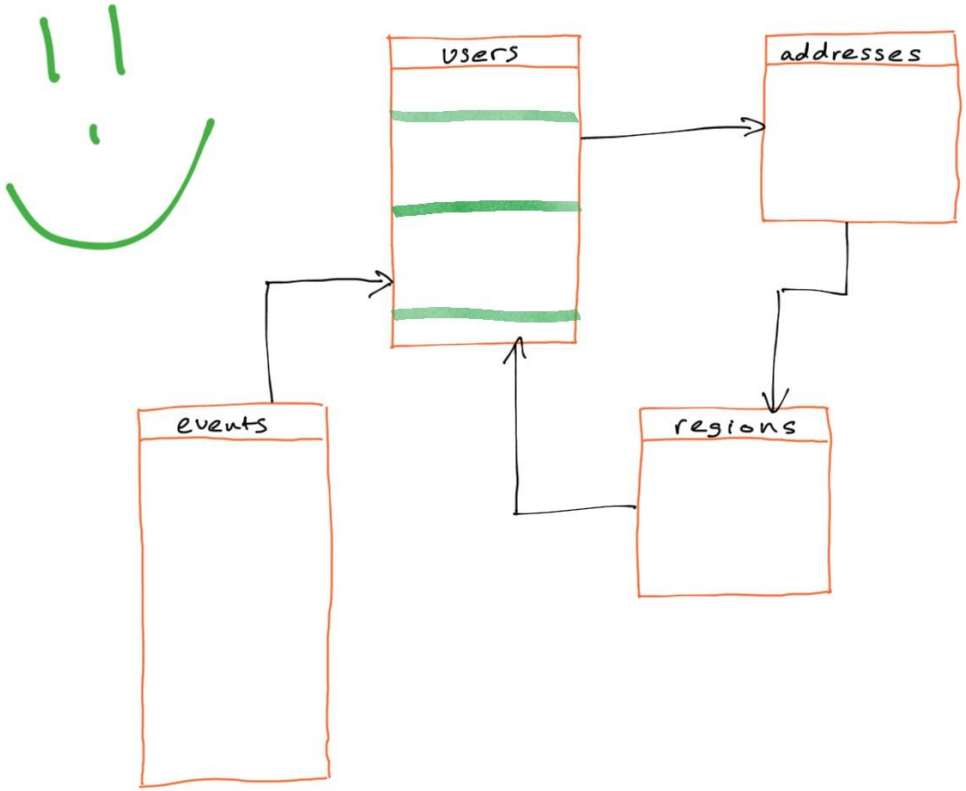
A Real DB



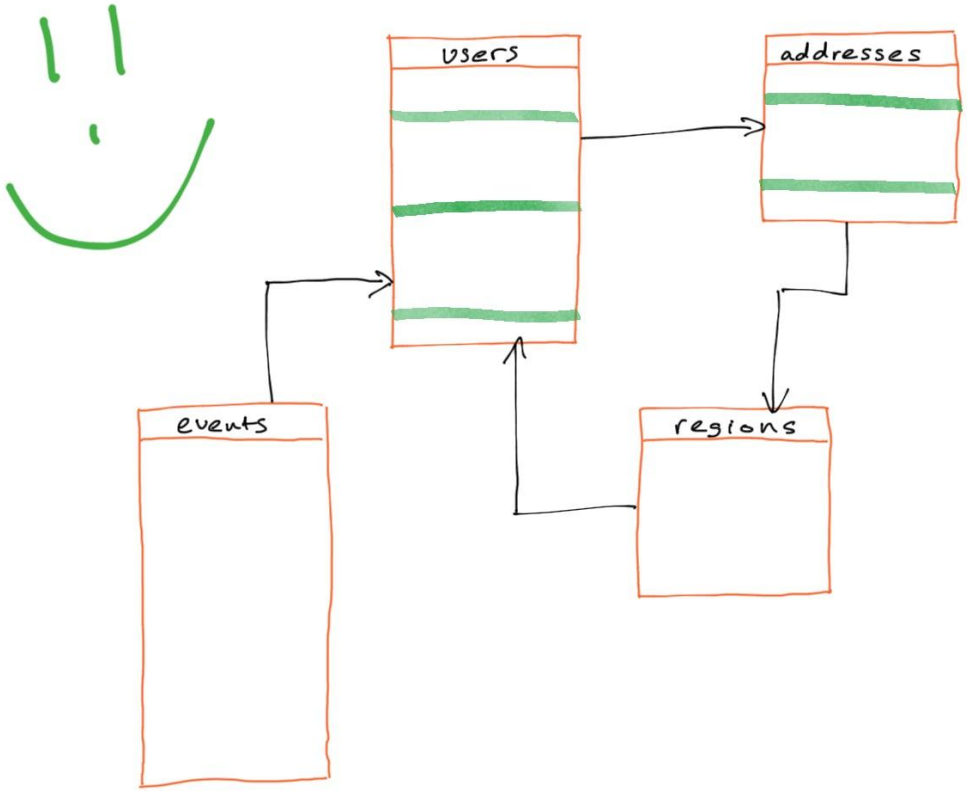
The Cycle Problem



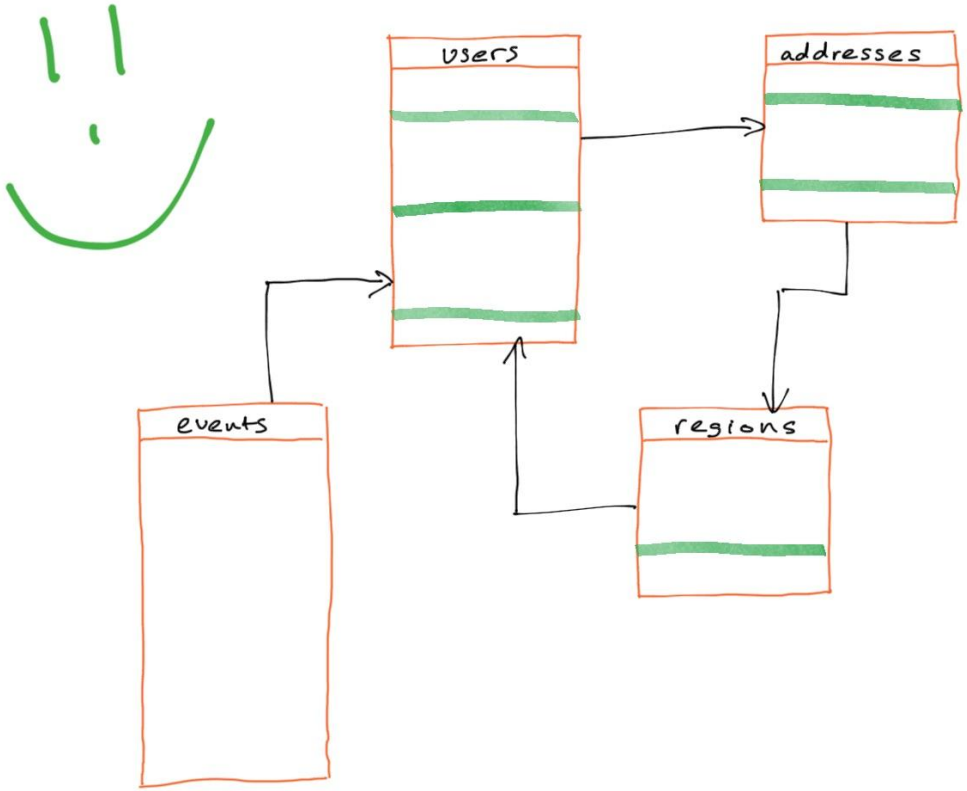
The Cycle Problem



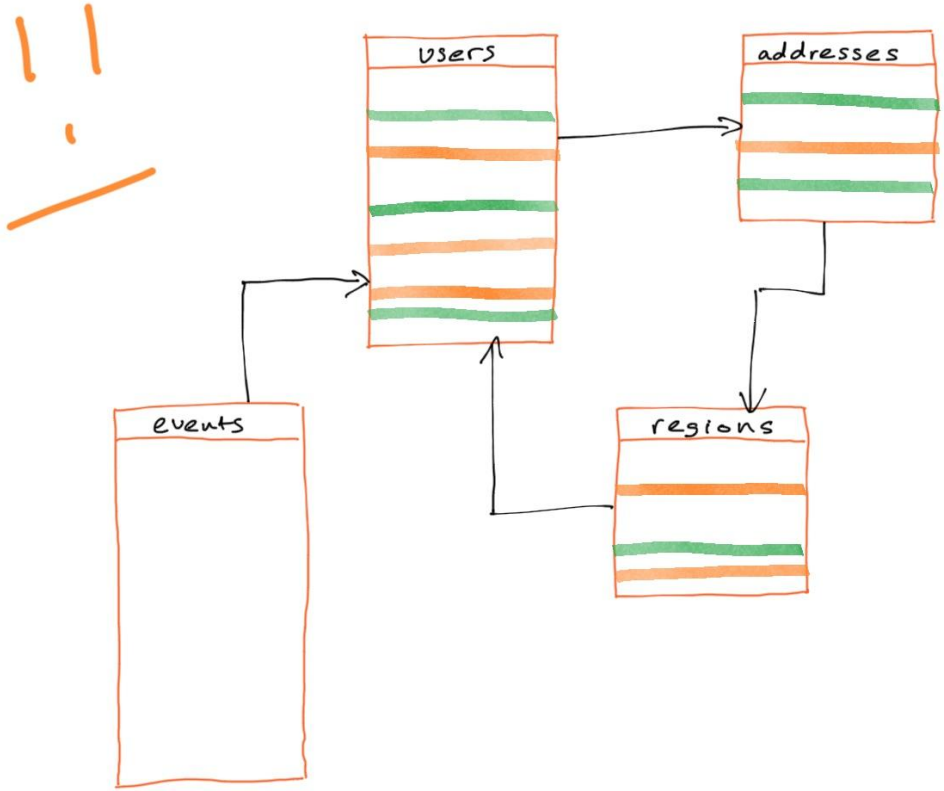
The Cycle Problem



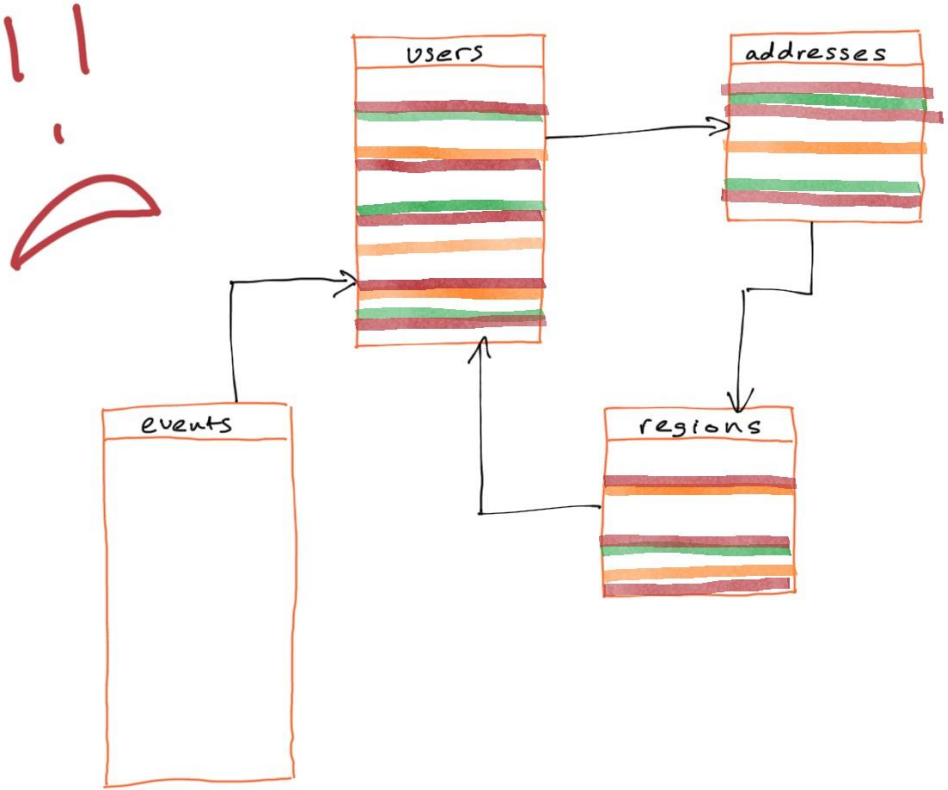
The Cycle Problem



The Cycle Problem



The Cycle Problem



The Cycle Problem

Cycles in databases are real. Examples.

- Social graphs have cycles
- Any time a table has a foreign key on itself
  - E.g. Event table can have a self-cycle to support previous event
- Backward key relations double the likelihood for cycles!

The (old) Cycle Solution

Expert configuration.

- Ignore certain tables/columns.
- Handle backward key relations in a custom way.

Not suitable for automatic subsetting by novice users.

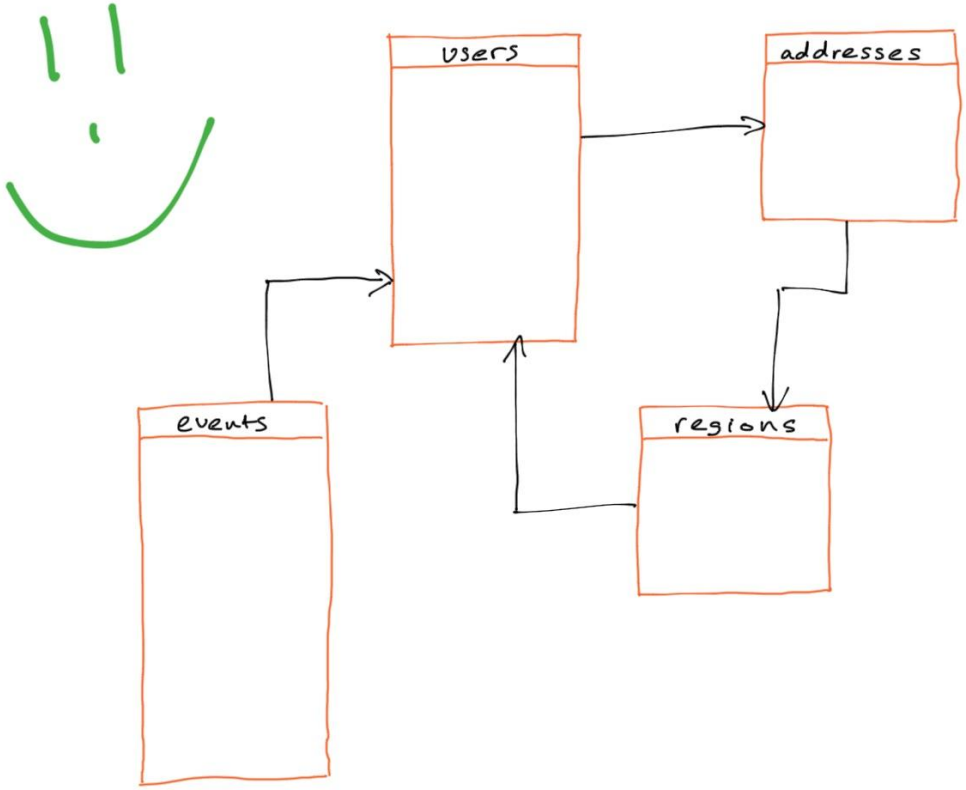
An expert in your DB

* Check out Jailer if you want to see an example of this. (Google “jailer subsetting”).

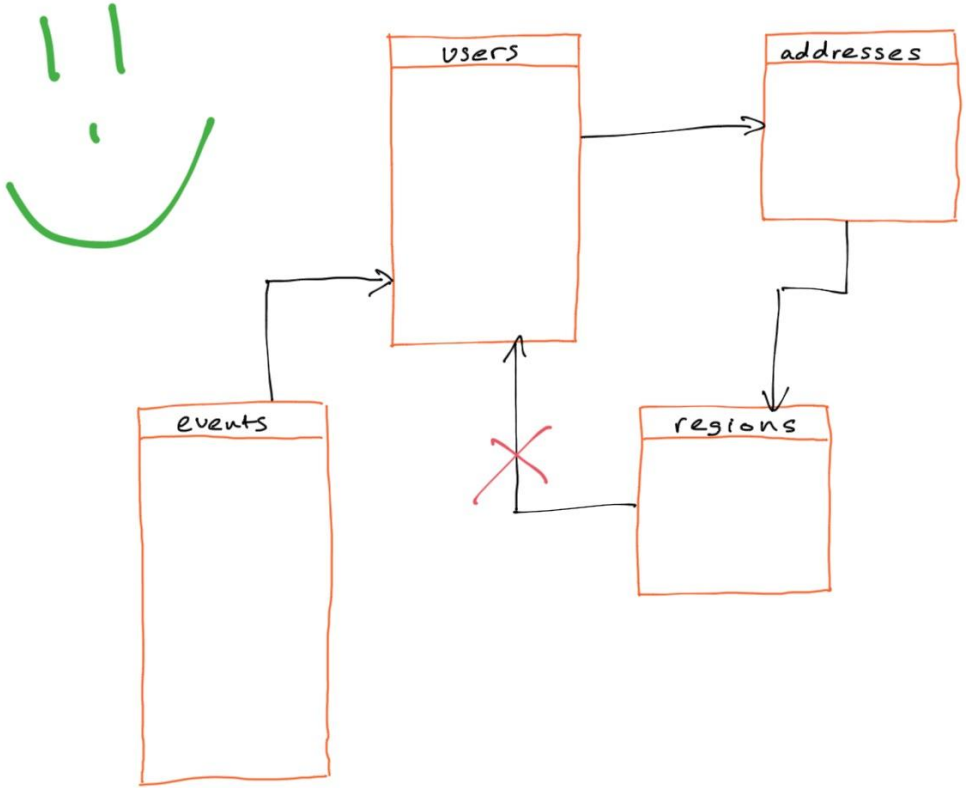
Our Cycle Solution

1. Cut the cycle(s).
2. Sort the tables by their dependency graph (topological sort).
    a. You can’t do topological sort if a cycle remains.
3. Feed the first table in the sort, push data through until you reach your target.

1. Cutting the Cycle



1. Cutting the Cycle

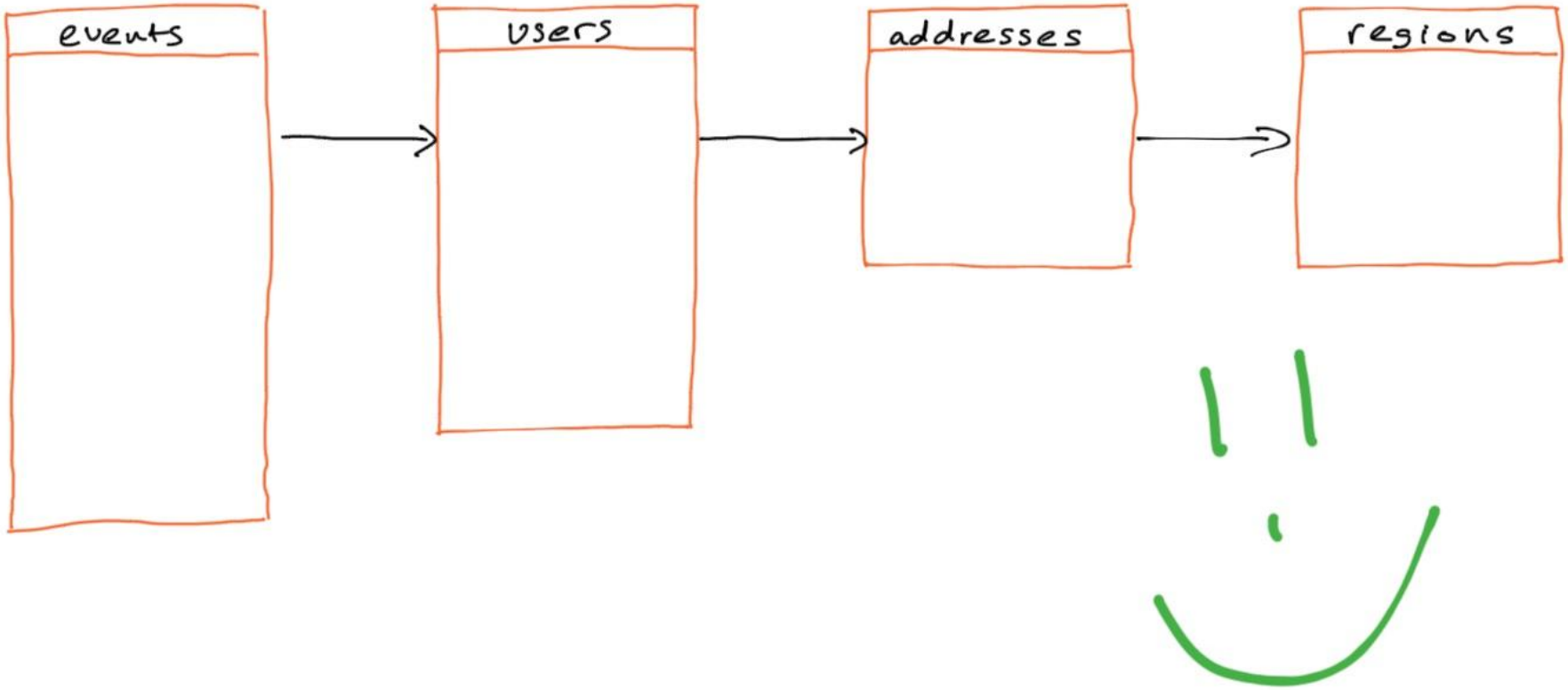


1. Cutting the Cycle

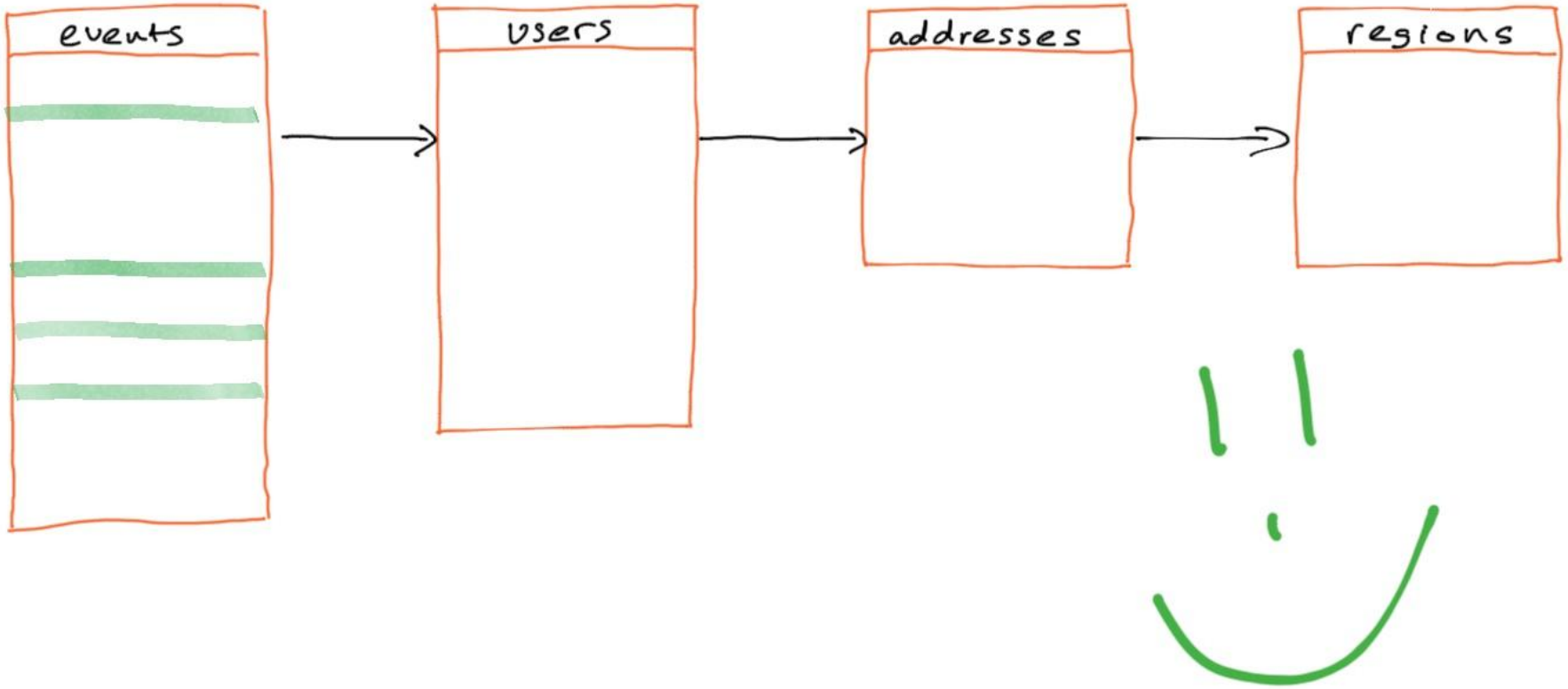
What does “cutting the cycle” mean?

- It means replacing the values of the foreign key in **regions** with **NULL**.
- *Problem:* What if schema doesn’t allow that? (I.e. NON NULL constraint.)
- *Solution:* There must be at least one NULLABLE column in any cycle.

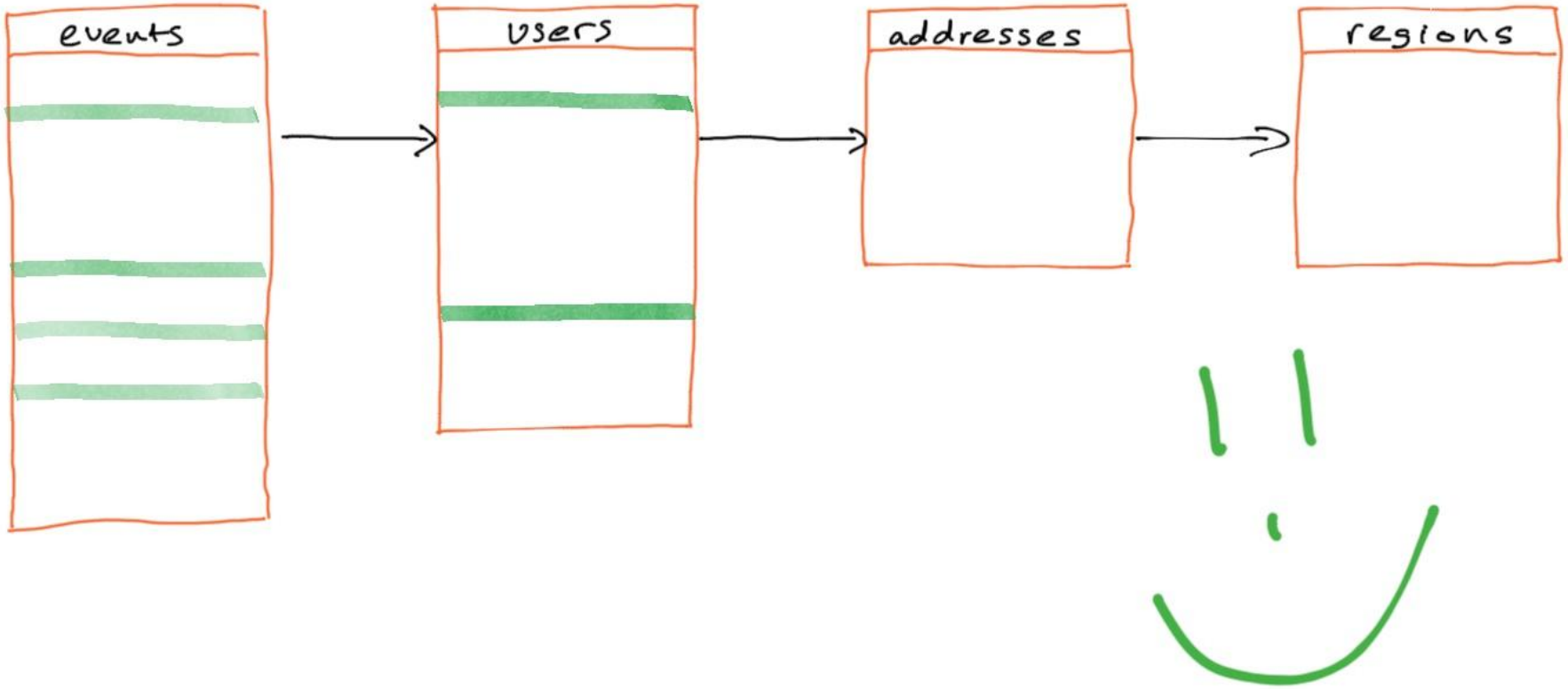
2. Topological Sort



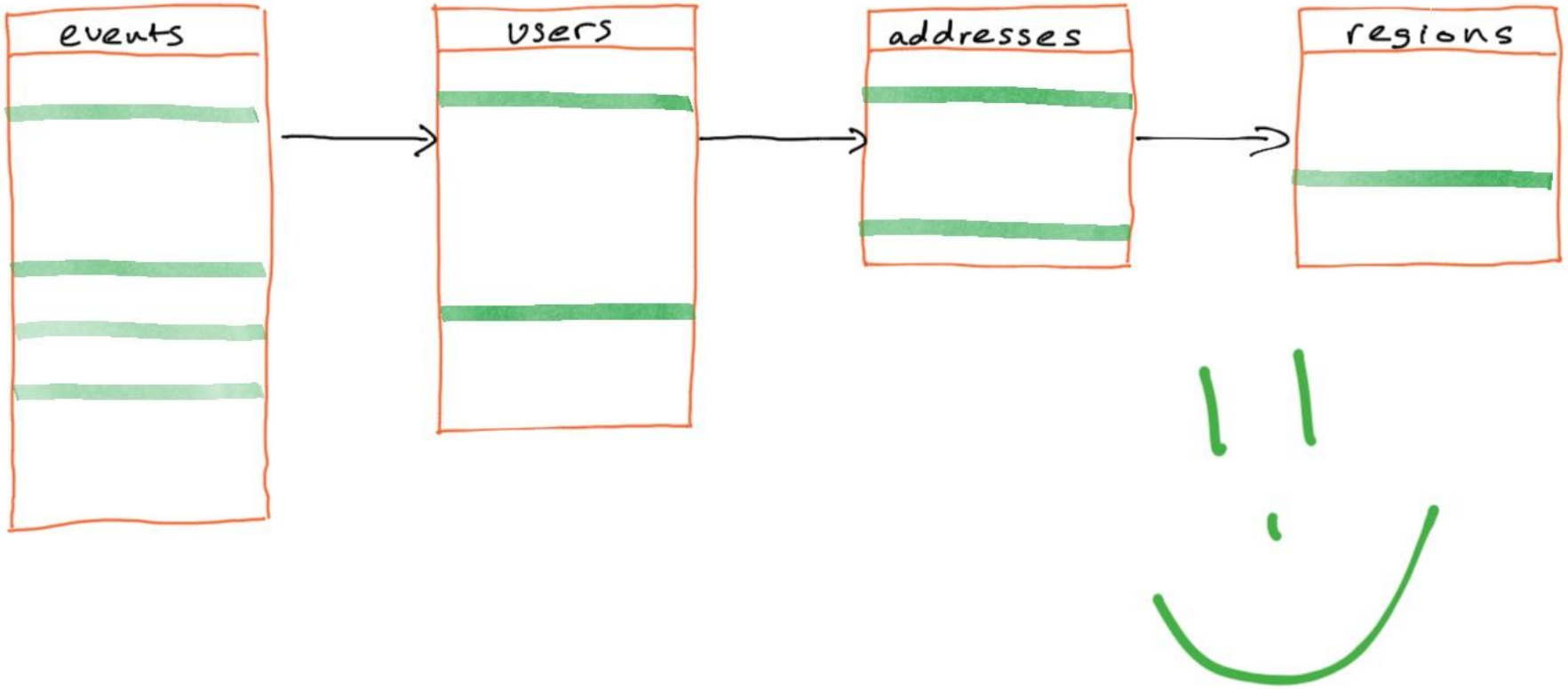
3. Feed the First, and Push Through



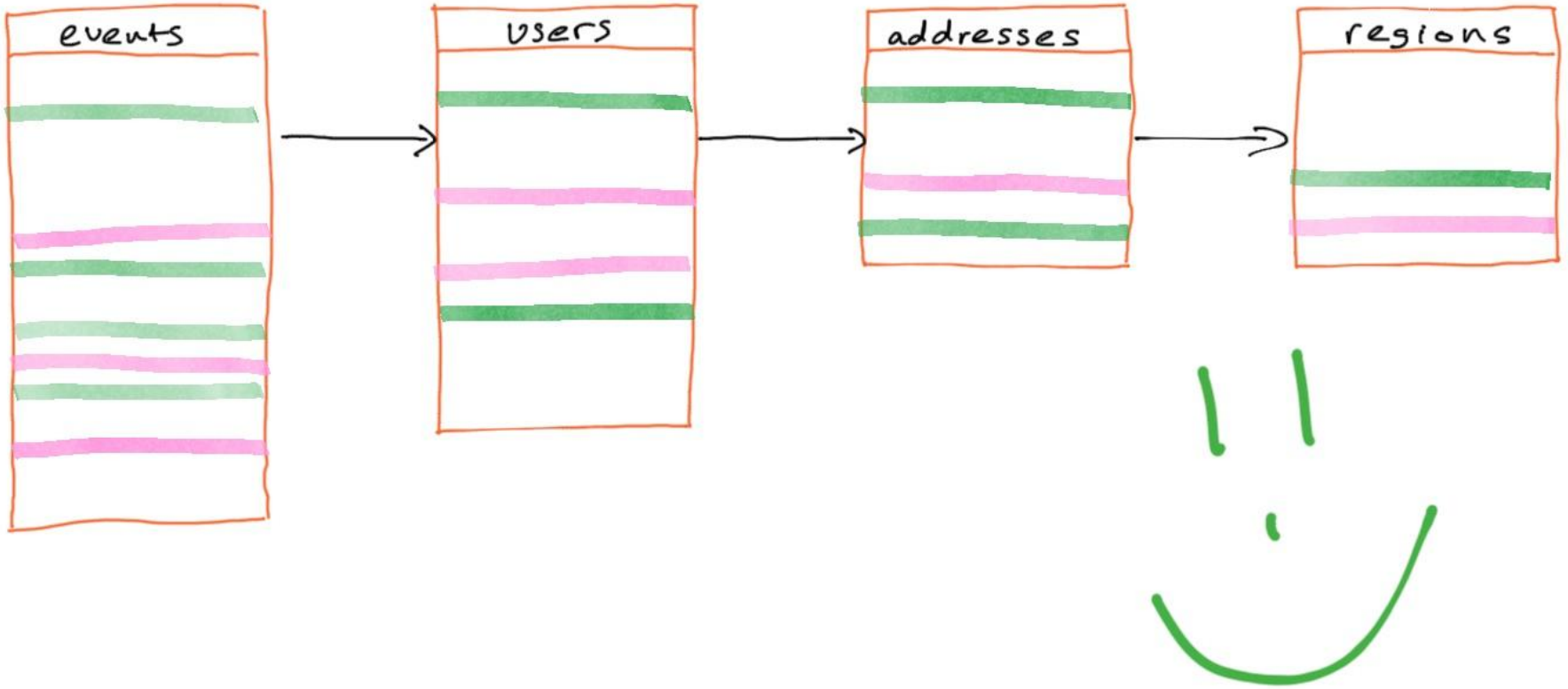
3. Feed the First, and Push Through



3. Feed the First, and Push Through



3. Feed the First, and Push Through



Outcomes

1. We didn’t accidentally take the whole database!
2. Configuration is simply the cycles to cut and our subset target.

Bonus outcome

3. Subset target can be any formula over the Database, E.g.
    - 5% of the whole DB.

Try it out

https://github.com/TonicAI/condenser