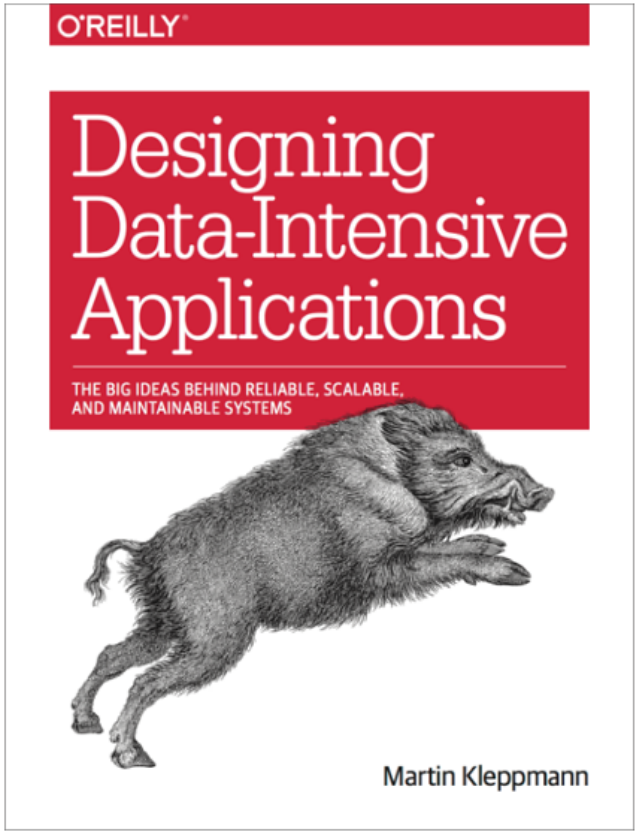
Building real-time

DATA PRODUCTS

samza

LinkedIn

dataintensive.net



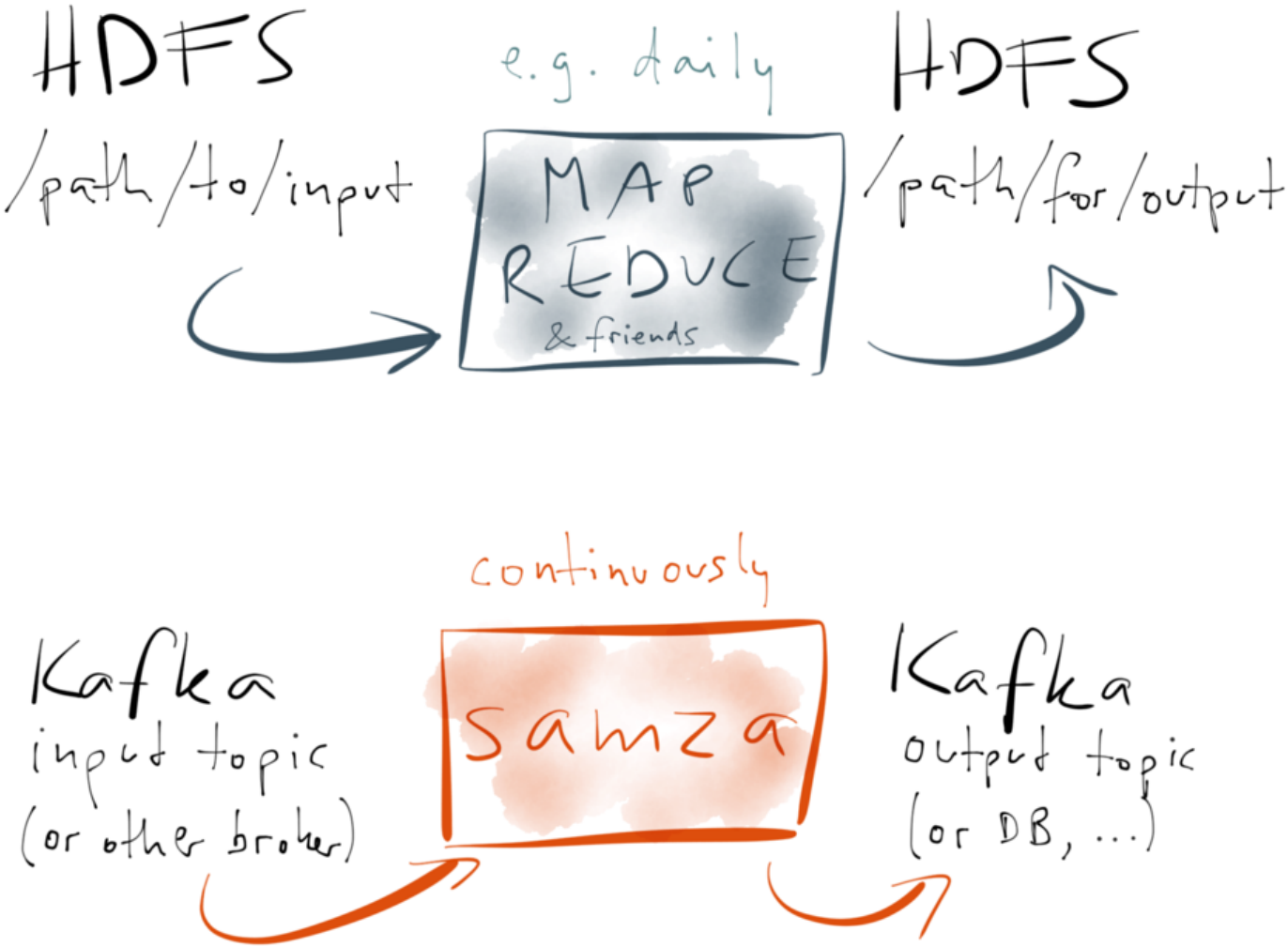
@martinkl

samza.incubator.apache.org

samza

DISTRIBUTED

STREAM PROCESSING


HDFS
e.g. daily
HDFS
/path/to/input
MAP
REDUCE
& friends
/path/for/output
continuously
Kafka
input topic
(or other broker)
samza
Kafka
output topic
(or DB, ...)

message broker

durable, fault-tolerant, high throughput, ordered stream, multi-subscriber

processing framework

reliable filtering, aggregation & joining of streams. deployment tooling.

YARN

samza

tez

MR

. . .

samza

etc . . .

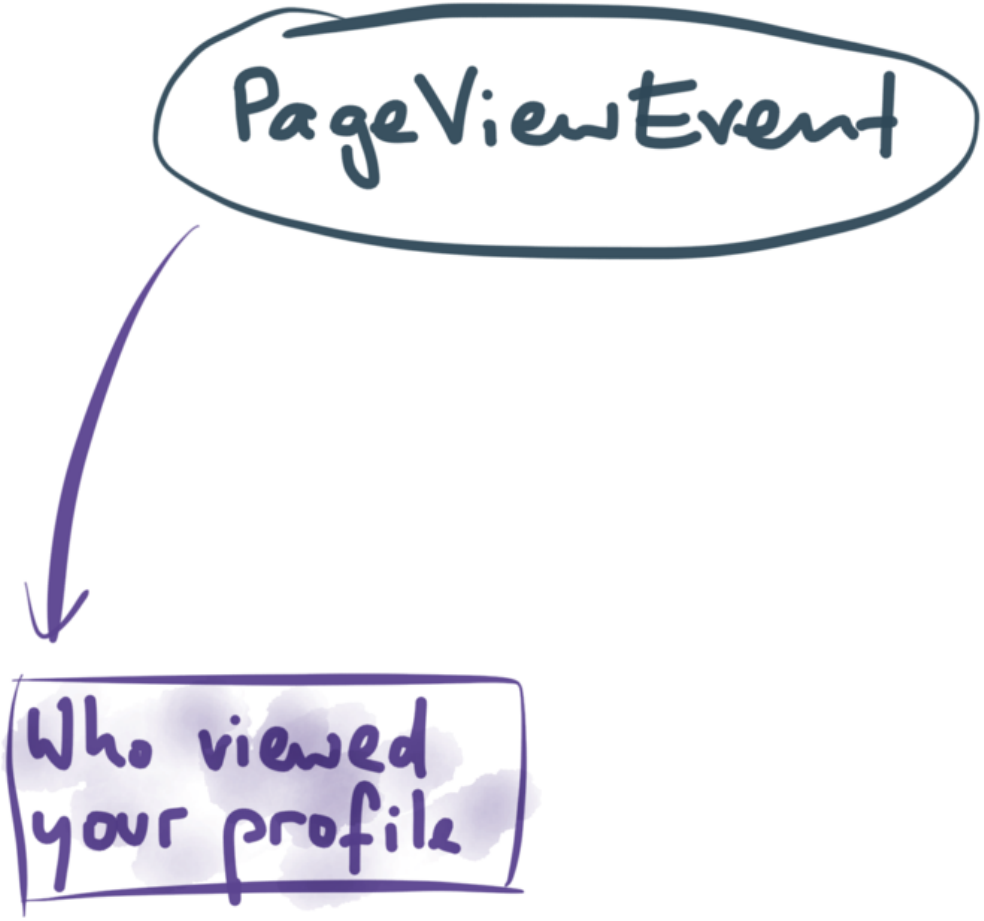
HDFS

Kafka

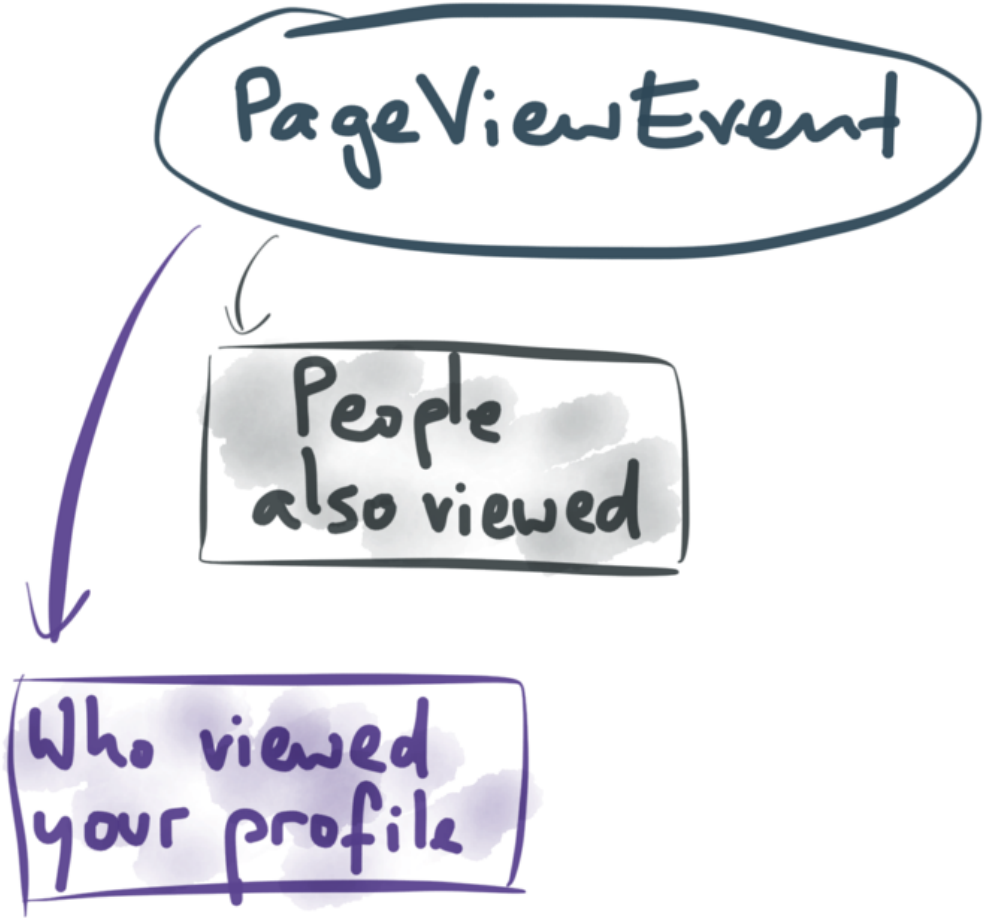
Enriching tracking events


PageViewEvent

```
{
    eventType:        PageViewEvent,
    timestamp:        1413215518,
    viewerId:         1234,
    sessionId:        fa1afe101234deadbeef,
    pageKey:          profile-view,
    viewedProfileId: 4321,
    trackingKey:      invitation-email,
    ... etc. metadata about what content was displayed...
}
```


PageViewEvent
Who viewed your profile


Search for people, jobs, companies, and more...
Advanced
Home
Profile
Network
Jobs
Interests
Business Tools
Upgrade
Top-Ranked MBA in SFO - Rated Top 1% by the Financial Times. Download your MBA eBrochure here! — Read More
Who's viewed your profile
How you rank for profile views
Hide charts
93 Profile views
Last 90 days
13 Viewers from the San Francisco Bay Area
7 Viewers from the Design Industry
9 Viewers found you from Google
Profile views
Actions taken
20
16
12
8
4
0
8
0
Dec 28
Jan 11
Jan 25
Feb 8
Feb 22
Mar 8
Mar 19
Jan 11 – Jan 18
16 Profile views
11% weekly change
6 Actions taken
24% weekly change
Actions taken include:
• 2 new connections
• Liked status update
• Endorsed 2 connections
• Posted status
Get more profile views
Viewers
James Brothers · 1st
Senior User Experience Designer at Microsoft
2h ago · Homepage
Message
Liana Tsai · 1st
Data Scientist at ZipApp
12h ago · People You May Know
Message
Veronica Capa · 2nd
Staff Software Engineer at Six28 Group
1d ago · LinkedIn Search
Connect
Keni B · 1st
System Manager at IBM
3d ago · People Similar to You
Message


PageViewEvent
People also viewed
Who viewed your profile

People Also Viewed

Rahul Vohra
LinkedIn Intro. CEO of Rapportive. Computer Scientist, Gamer, Entrepreneur.

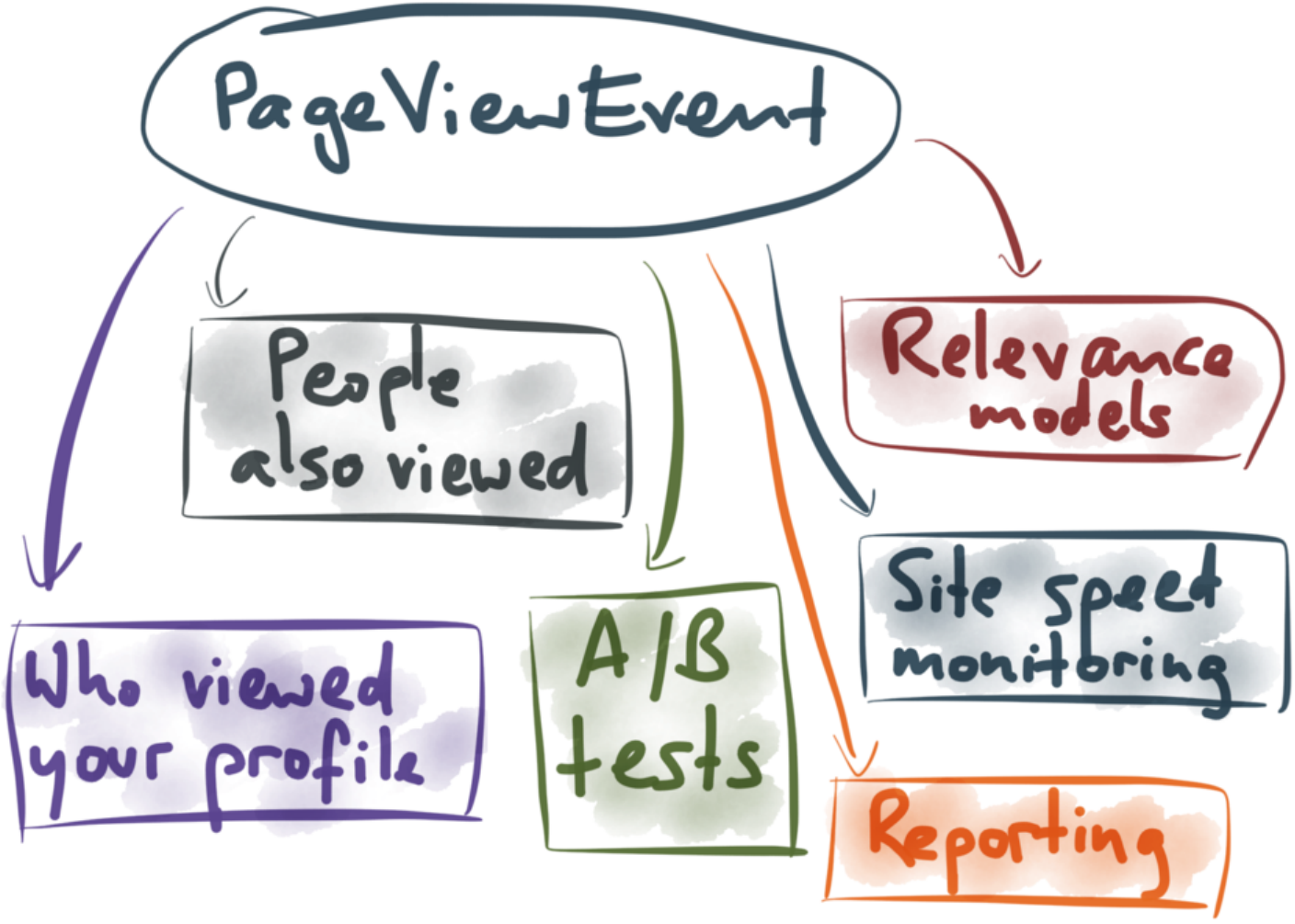
Sam Stokes
Builder, technologist, mixologist

Jay Kreps
Principal Staff Engineer at LinkedIn

Rashmi Sinha
Cofounder and Head at SlideShare, a LinkedIn company

Lee Mallabone
Senior Engineer at LinkedIn

Jakob Homan
Hadoop Ecosystem Engineer


PageViewEvent
People also viewed
Relevance models
Who viewed your profile
A/B tests
Site speed monitoring
Reporting

Who's viewed your profile | How you rank for profile views

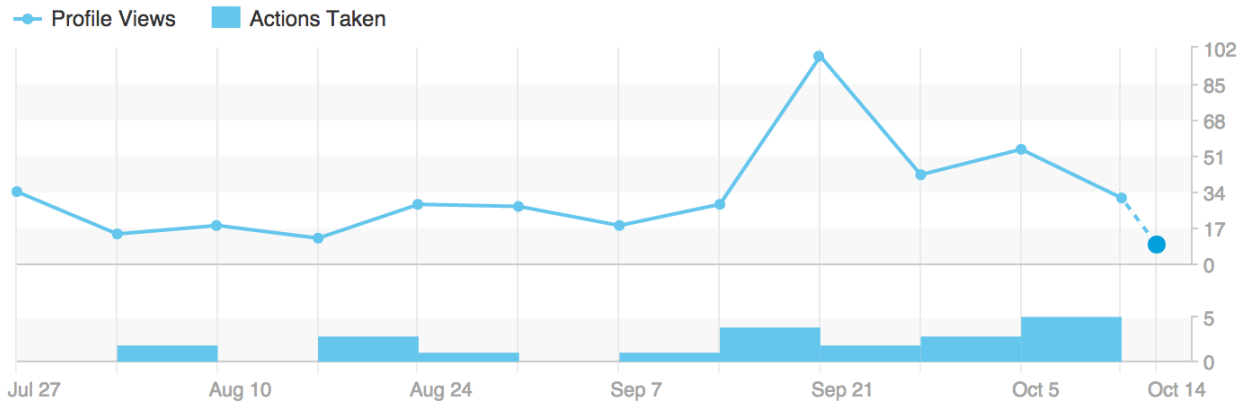
Hide charts

411 **Profile views**
Last 90 days

98 Viewers found you from People similar to you

27 Viewers who work at LinkedIn

70 Viewers with the title Software Developer



Oct 13 – now

9 **Profile views**
-- from the previous week

0 **Actions taken**
-- from the previous week

Getting more profile views can help you get found for the right opportunity.

Get more profile views

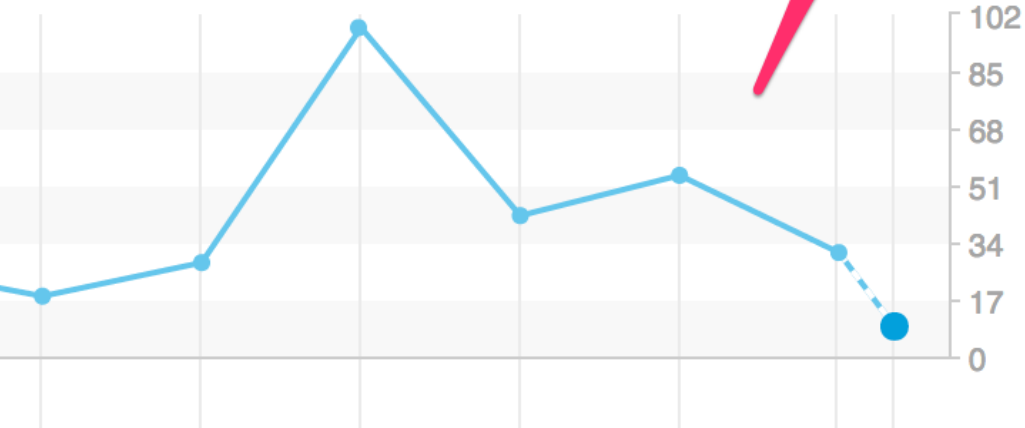
How you rank for profile views

Hide charts

s found you from
similar to you

27 Viewers who work at LinkedIn

70 Viewers with the title Software Developer



Oct 13 – now

9 **Profile views**
-- from the previous week

0 **Actions taken**
-- from the previous week

Getting more profile views can help you get found for the right opportunity.

{viewerId:1234}

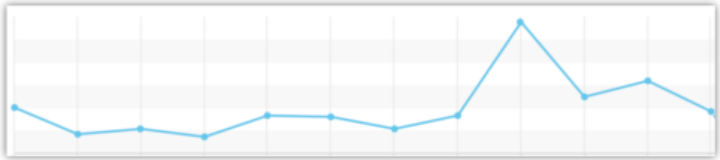
enrich with profile data

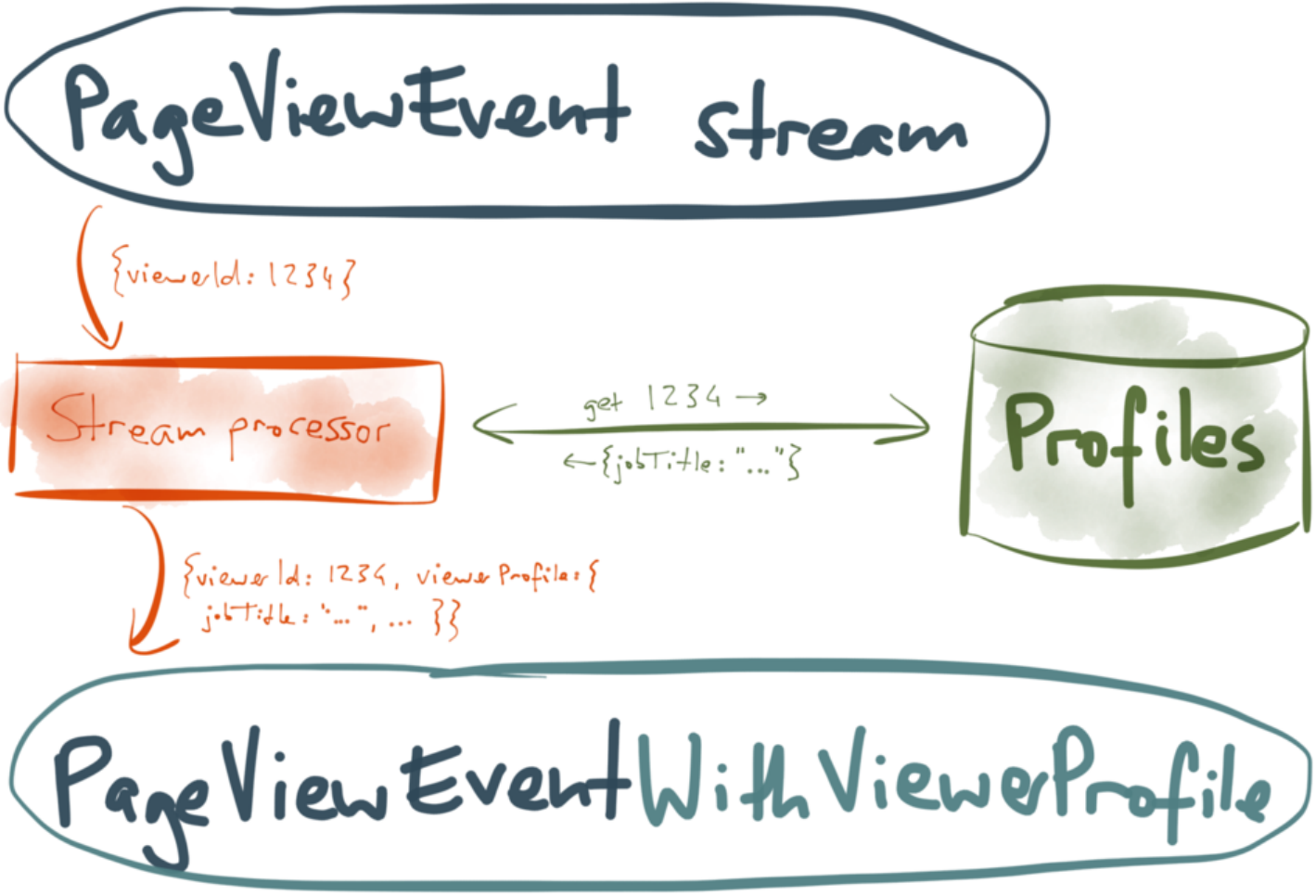
{viewerId: 1234, viewerProfile:{

jobTitle: "Software Engineer"}}

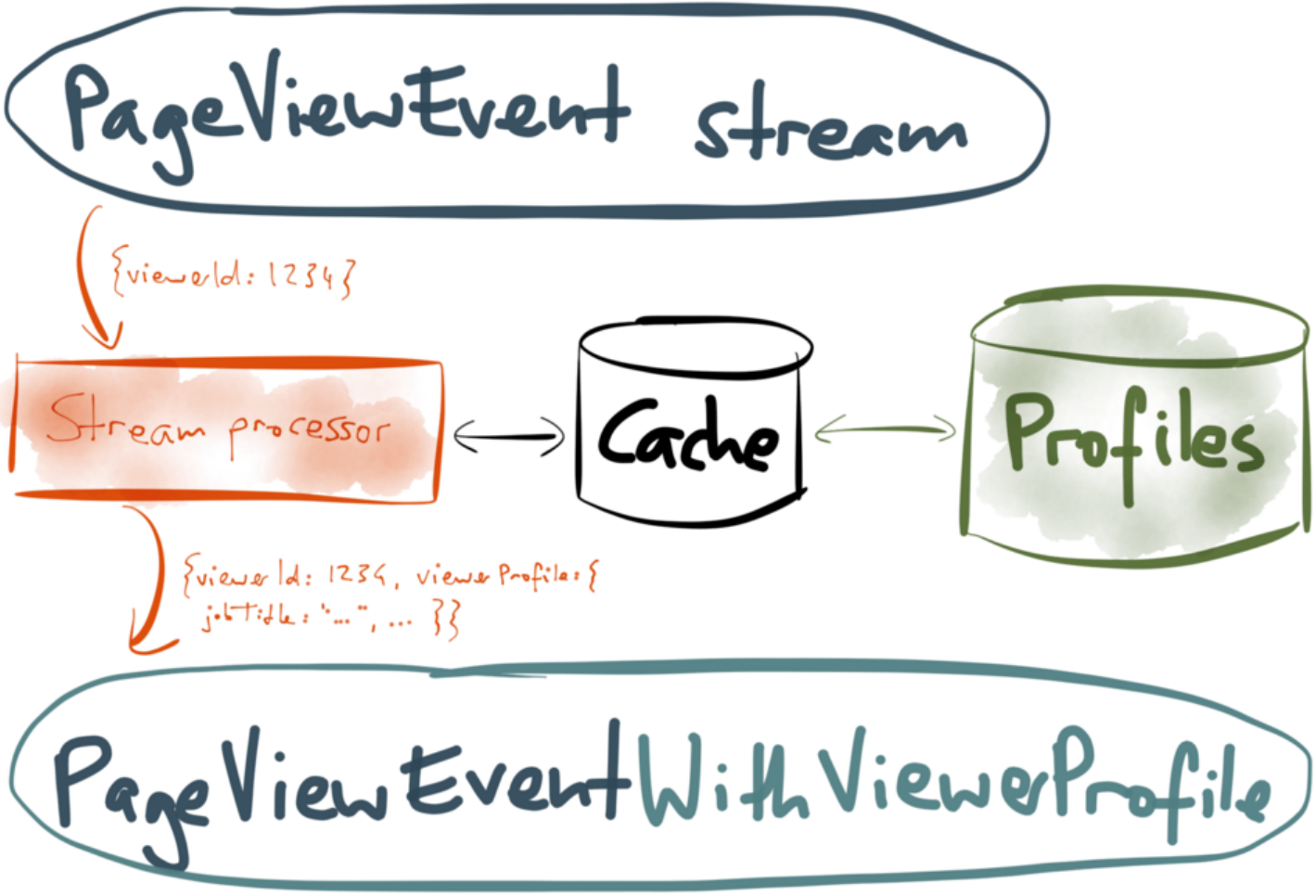
indexing

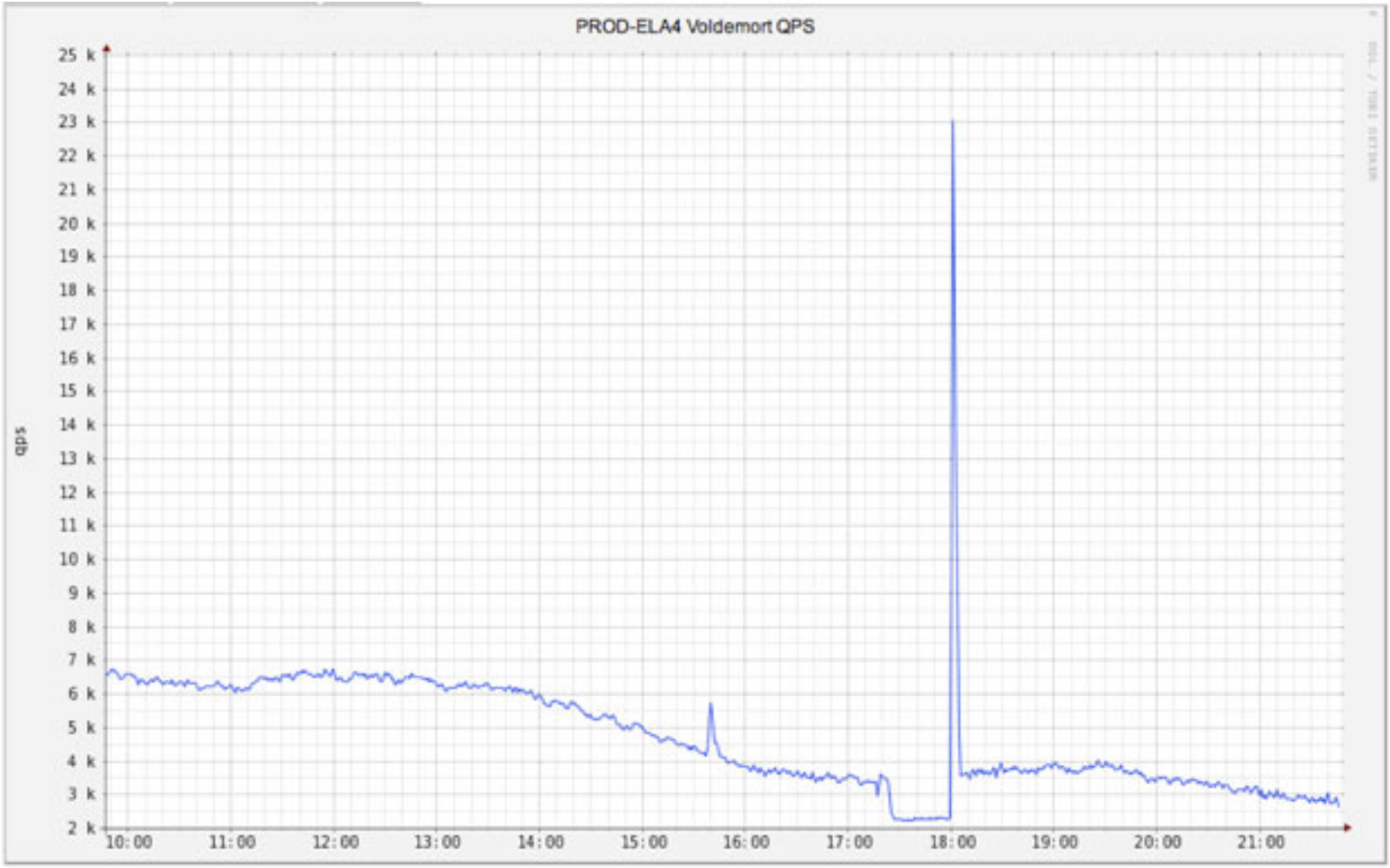
70 Viewers with the title Software Developer



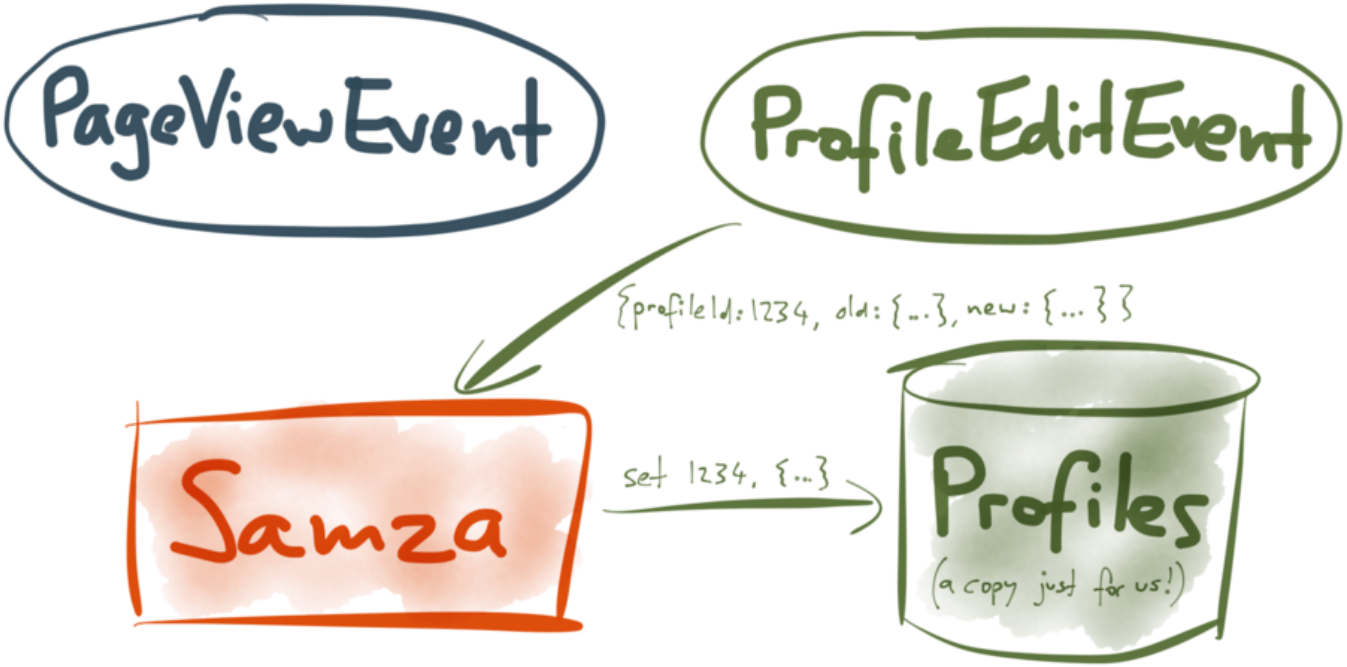

PageViewEvent stream
{viewerId: 1234}
Stream processor
get 1234 →
←{jobTitle: "..."}
Profiles
{viewerId: 1234, viewerProfile: {
jobTitle: "...", ... }}
PageViewEventWithViewerProfile

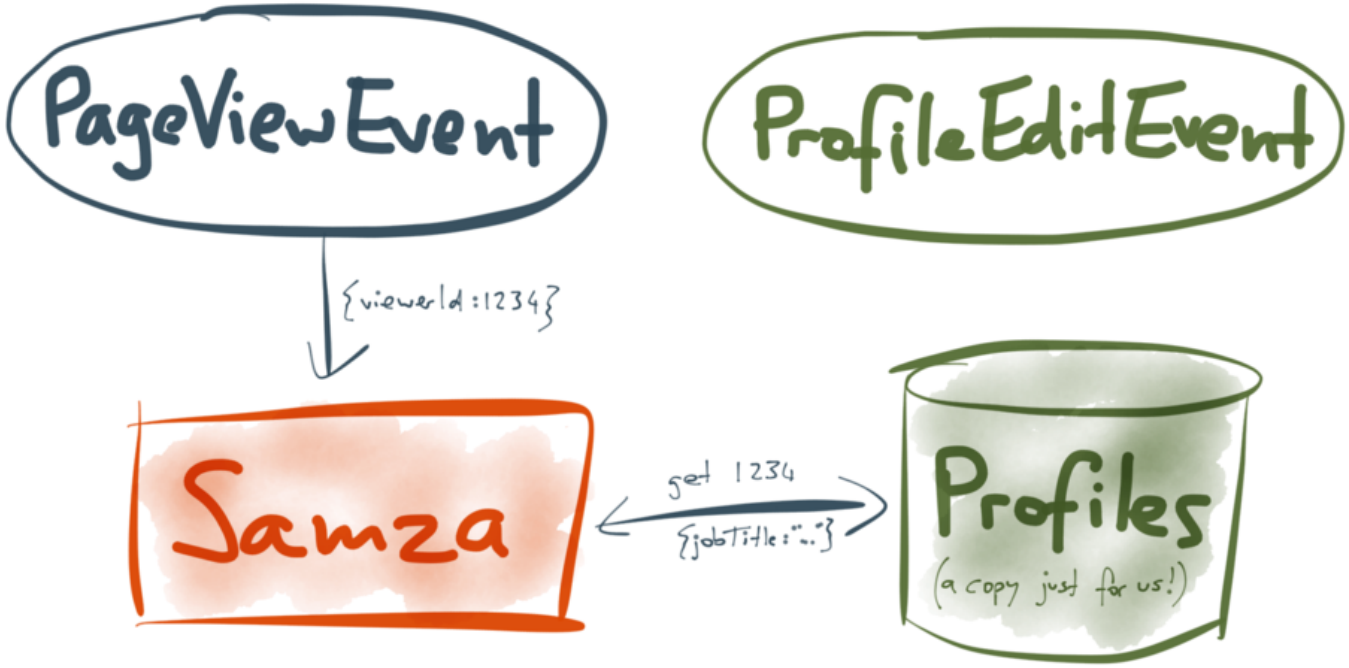
too slow!

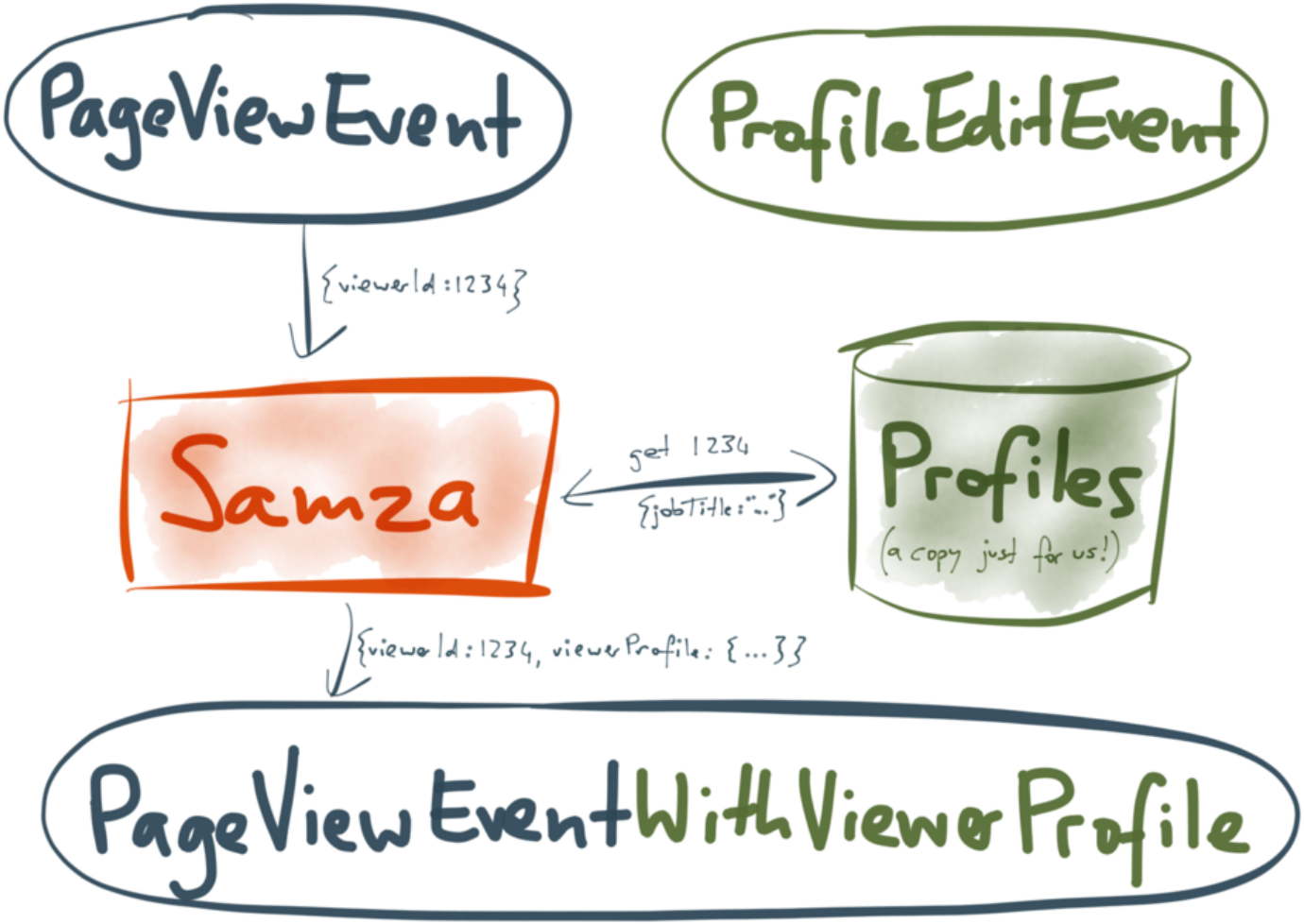

PageViewEvent stream
{viewId: 1234}
Stream processor
Cache
Profiles
{viewId: 1234, viewerProfile: {
jobTitle: "...", ... }}
PageViewEventWithViewerProfile

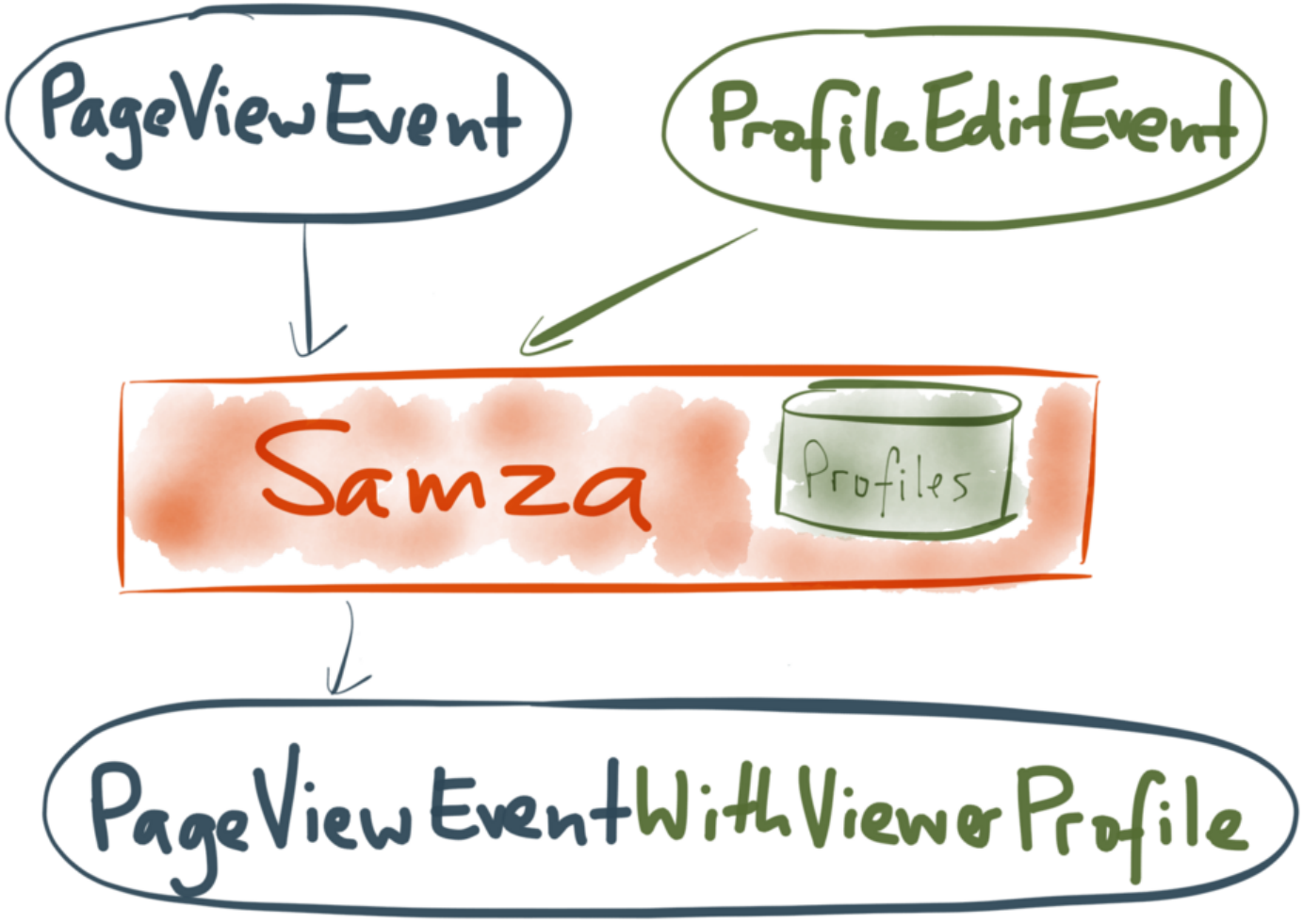

PROD-ELA4 Voldemort QPS
qps
25 k
24 k
23 k
22 k
21 k
20 k
19 k
18 k
17 k
16 k
15 k
14 k
13 k
12 k
11 k
10 k
9 k
8 k
7 k
6 k
5 k
4 k
3 k
2 k
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00

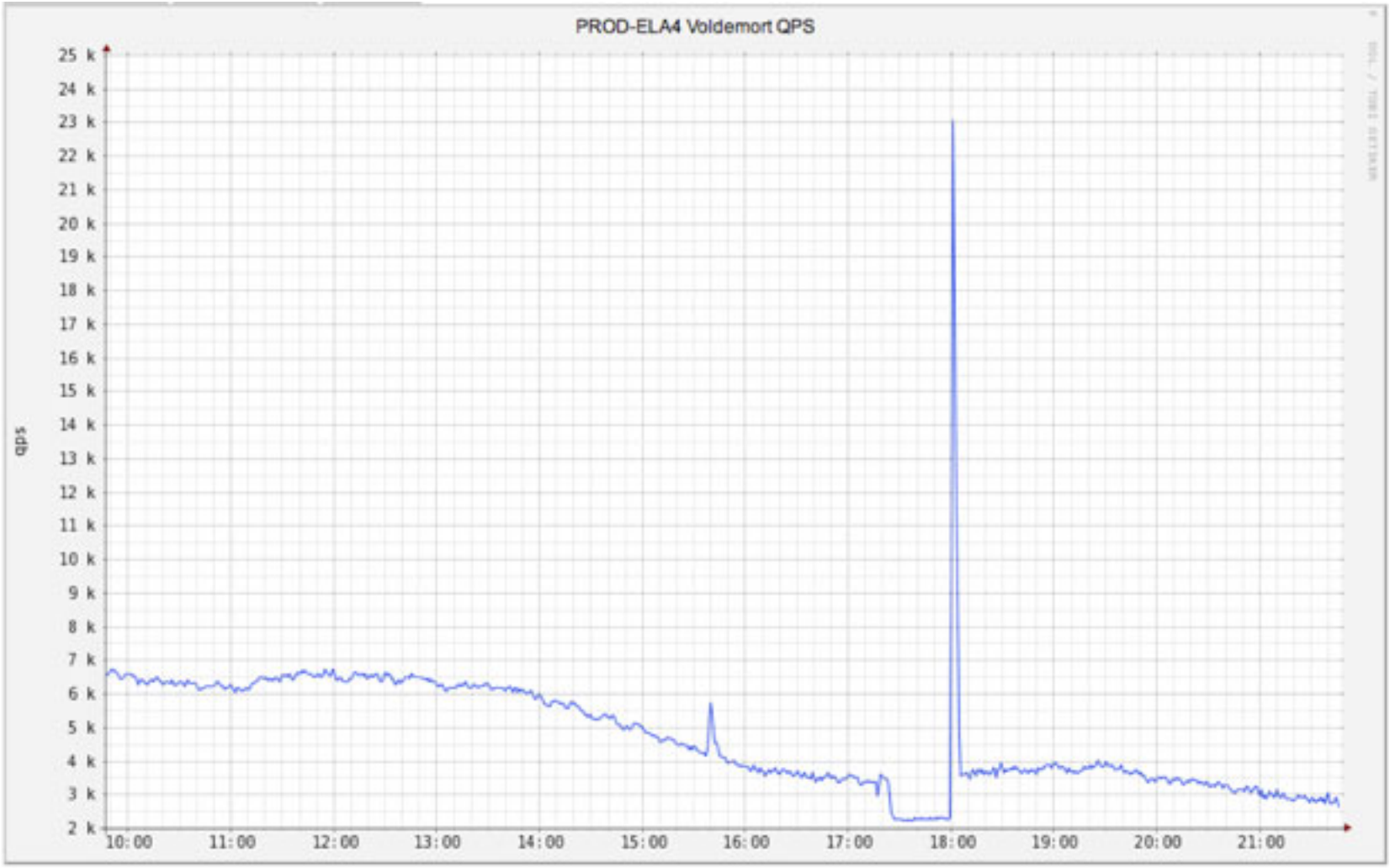
```
{
    eventType:      ProfileEditEvent,
    timestamp:      1413215518,
    profileId:      1234,
    old: {
        location:   "San Francisco, CA",
        industry:   "Internet"},
    new: {
        location:   "New York, NY",
        industry:   "Financial Services"}
}
```

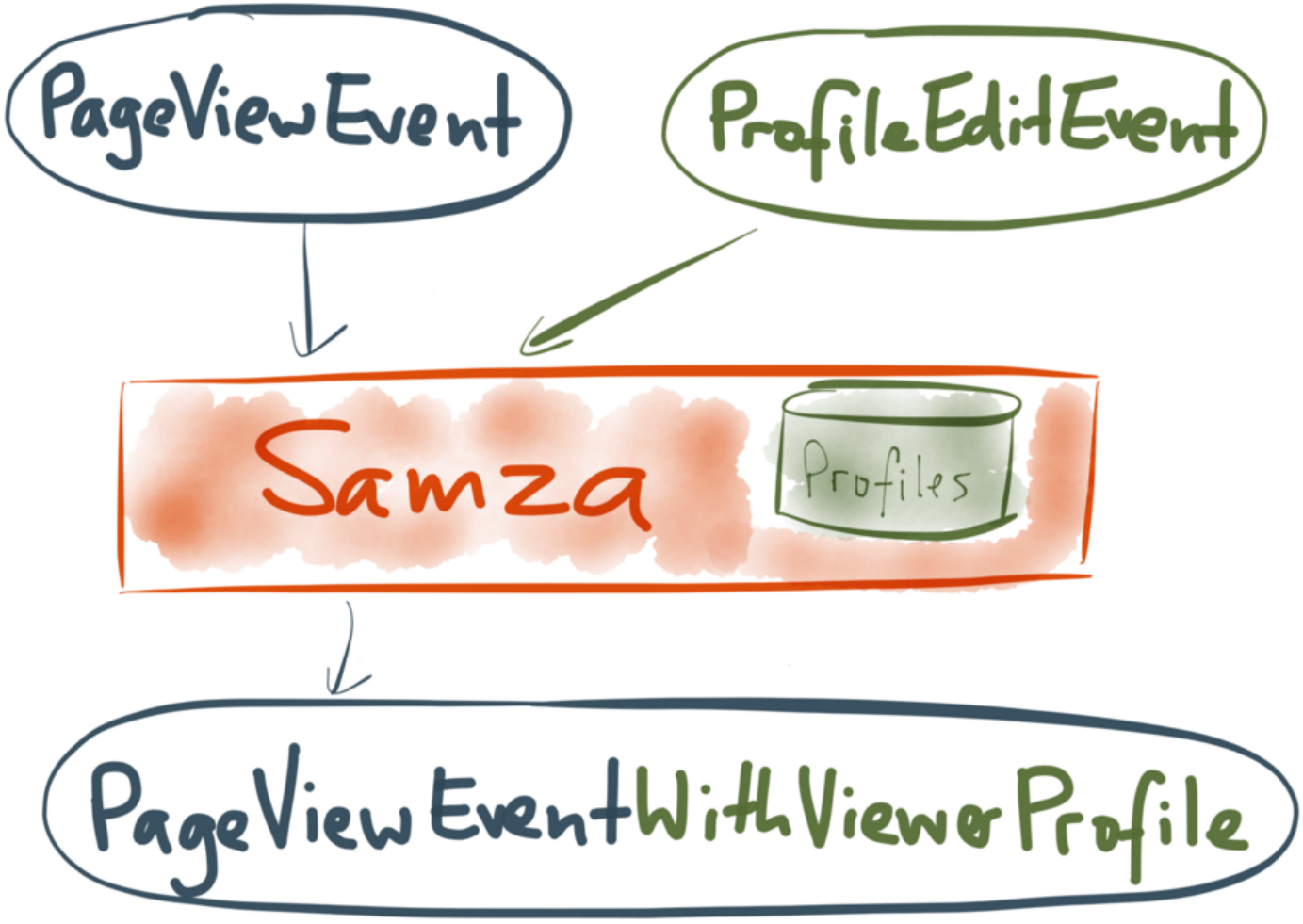

PageViewEvent
ProfileEditEvent
{profileId:1234, old:{...}, new:{...}}
Samza
set 1234, {...}
Profiles
(a copy just for us!)


PageViewEvent
ProfileEditEvent
{viewerId:1234}
Samza
get 1234
{jobTitle:"..."}
Profiles
(a copy just for us!)


PageViewEvent
ProfileEditEvent
{viewerId:1234}
Samza
get 1234
{jobTitle:"..."}
Profiles
(a copy just for us!)
{viewerId:1234, viewerProfile: {...}}
PageViewEventWithViewerProfile


PageViewEvent
ProfileEditEvent
Samza
Profiles
PageViewEventWithViewerProfile

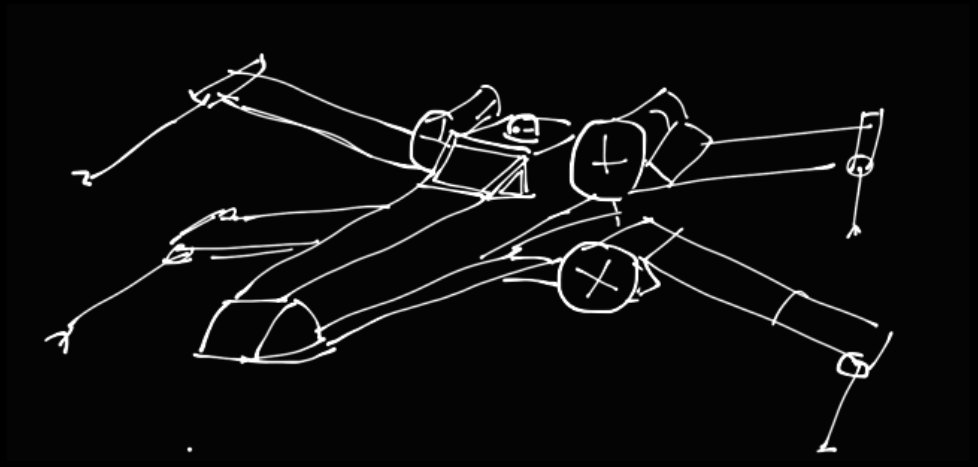

PROD-ELA4 Voldemort QPS
qps
25 k
24 k
23 k
22 k
21 k
20 k
19 k
18 k
17 k
16 k
15 k
14 k
13 k
12 k
11 k
10 k
9 k
8 k
7 k
6 k
5 k
4 k
3 k
2 k
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00


PageViewEvent
ProfileEditEvent
Samza
Profiles
PageViewEventWithViewerProfile

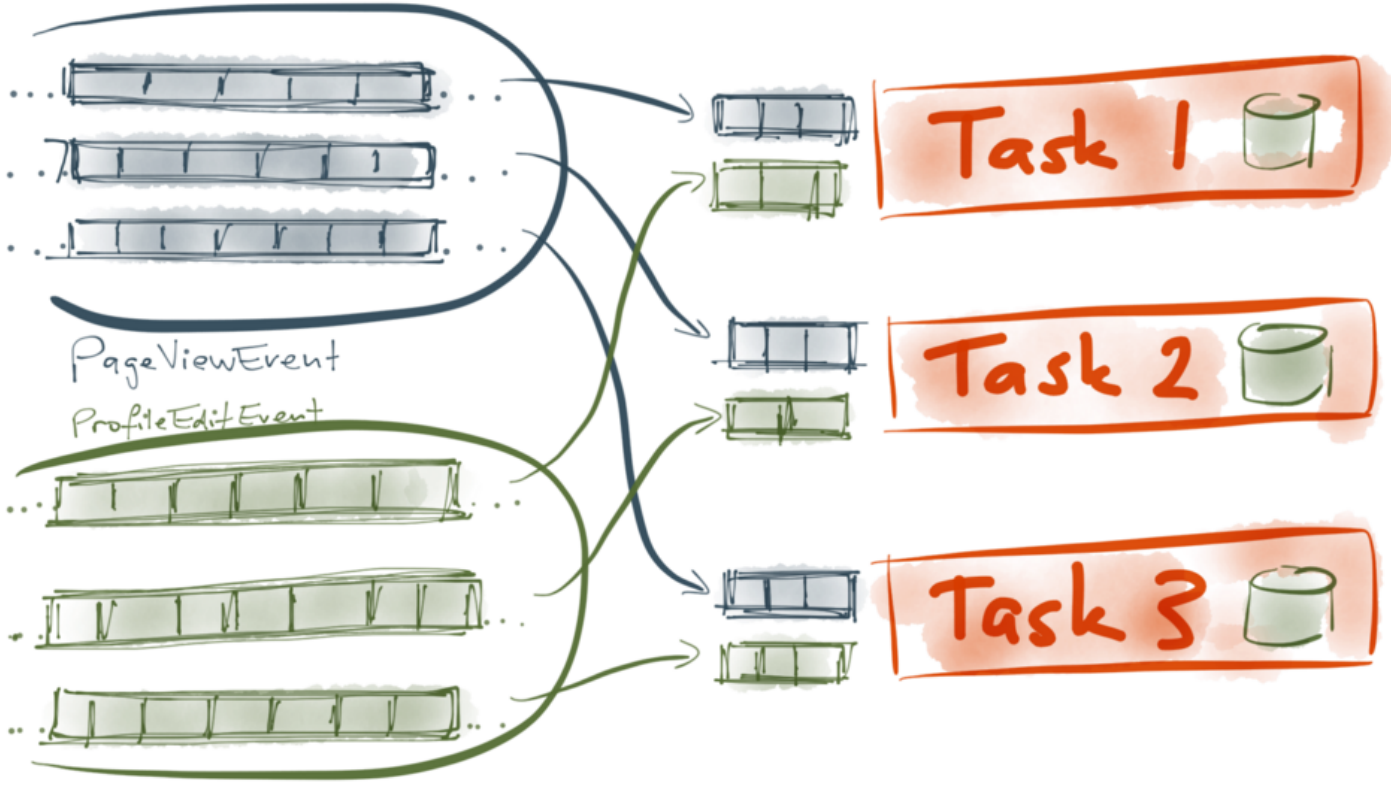

JOINZ
NOT DEAD

SAMZA

RETURN OF THE JOIN

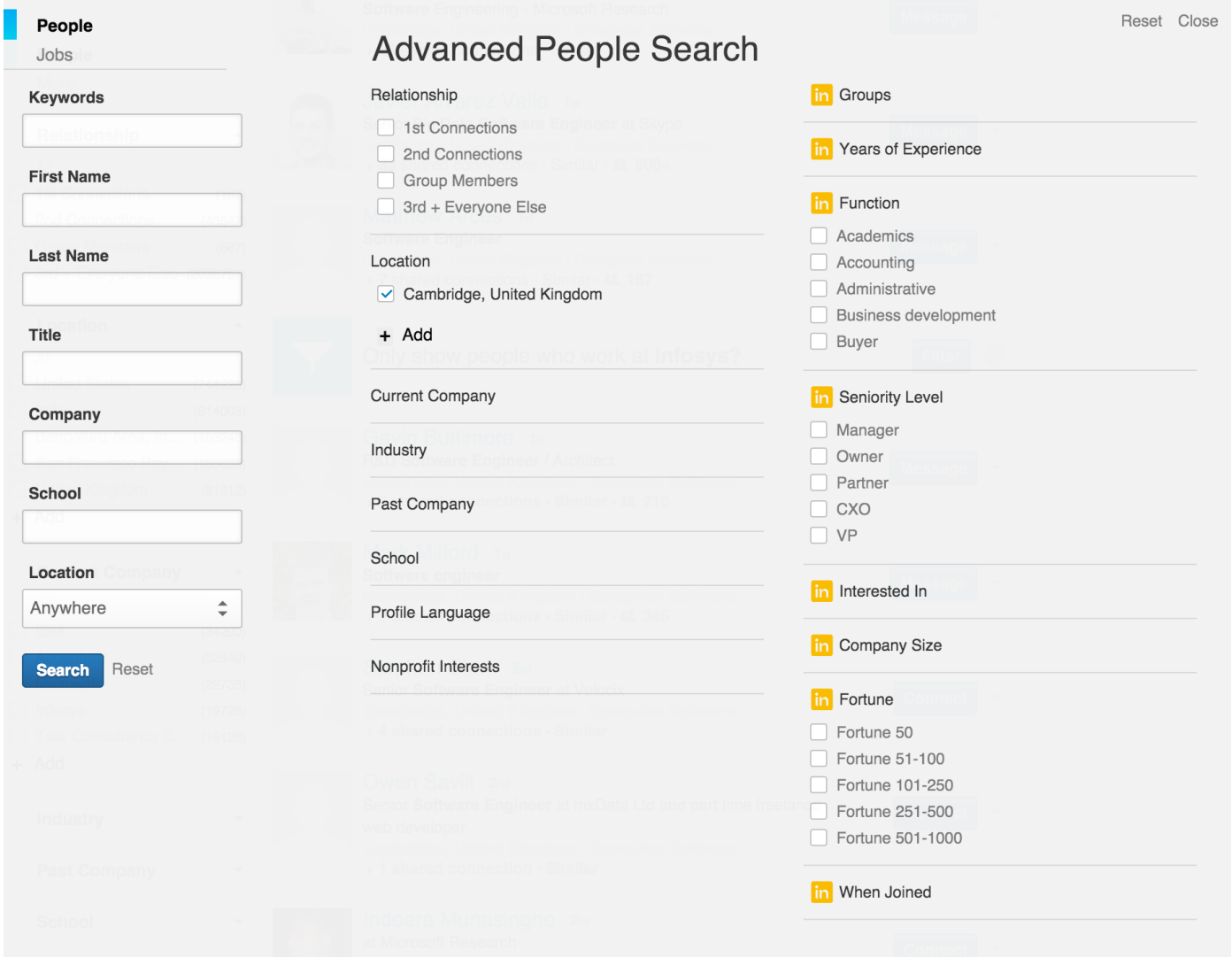


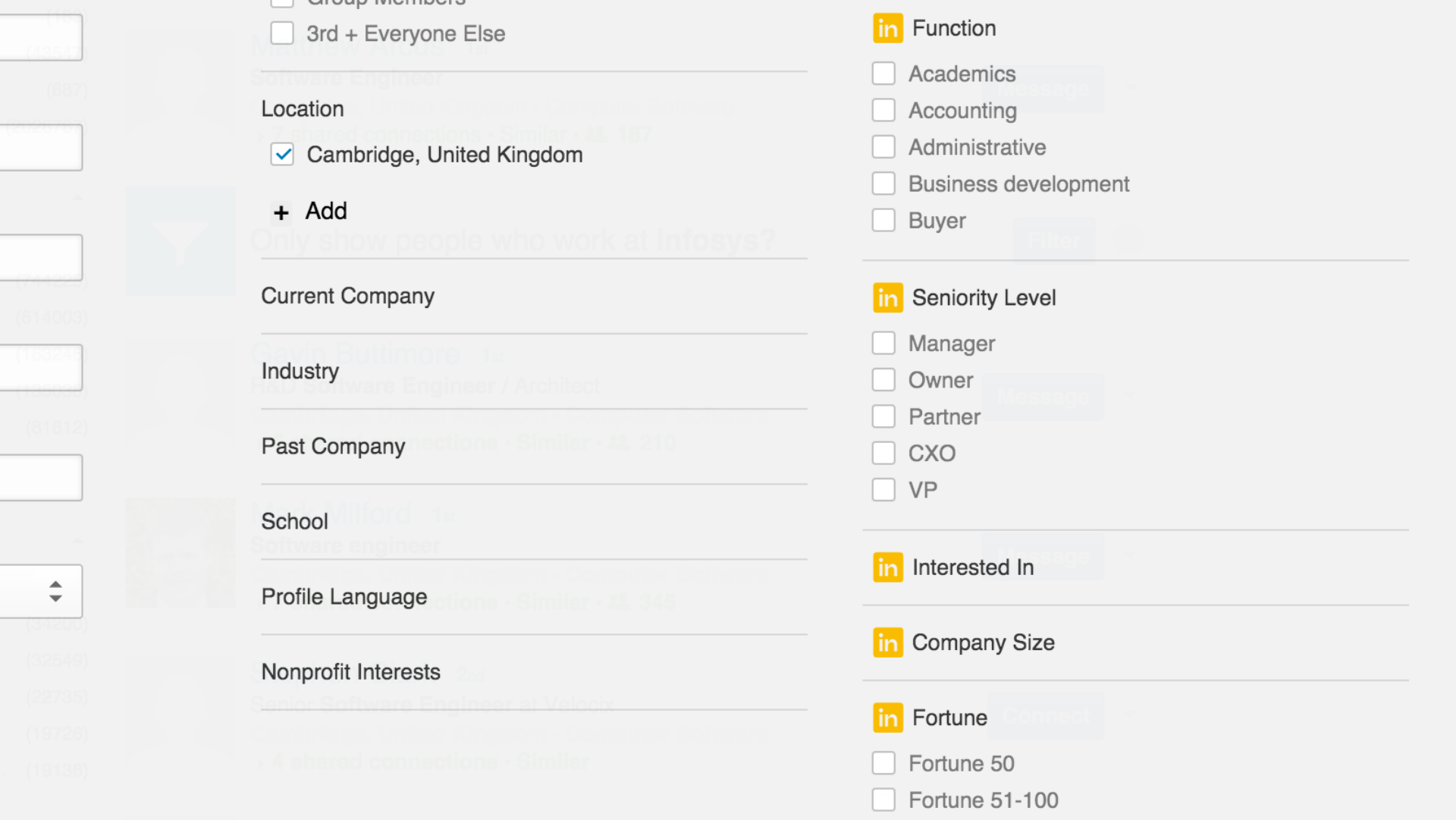
Scaling by co-partitioning

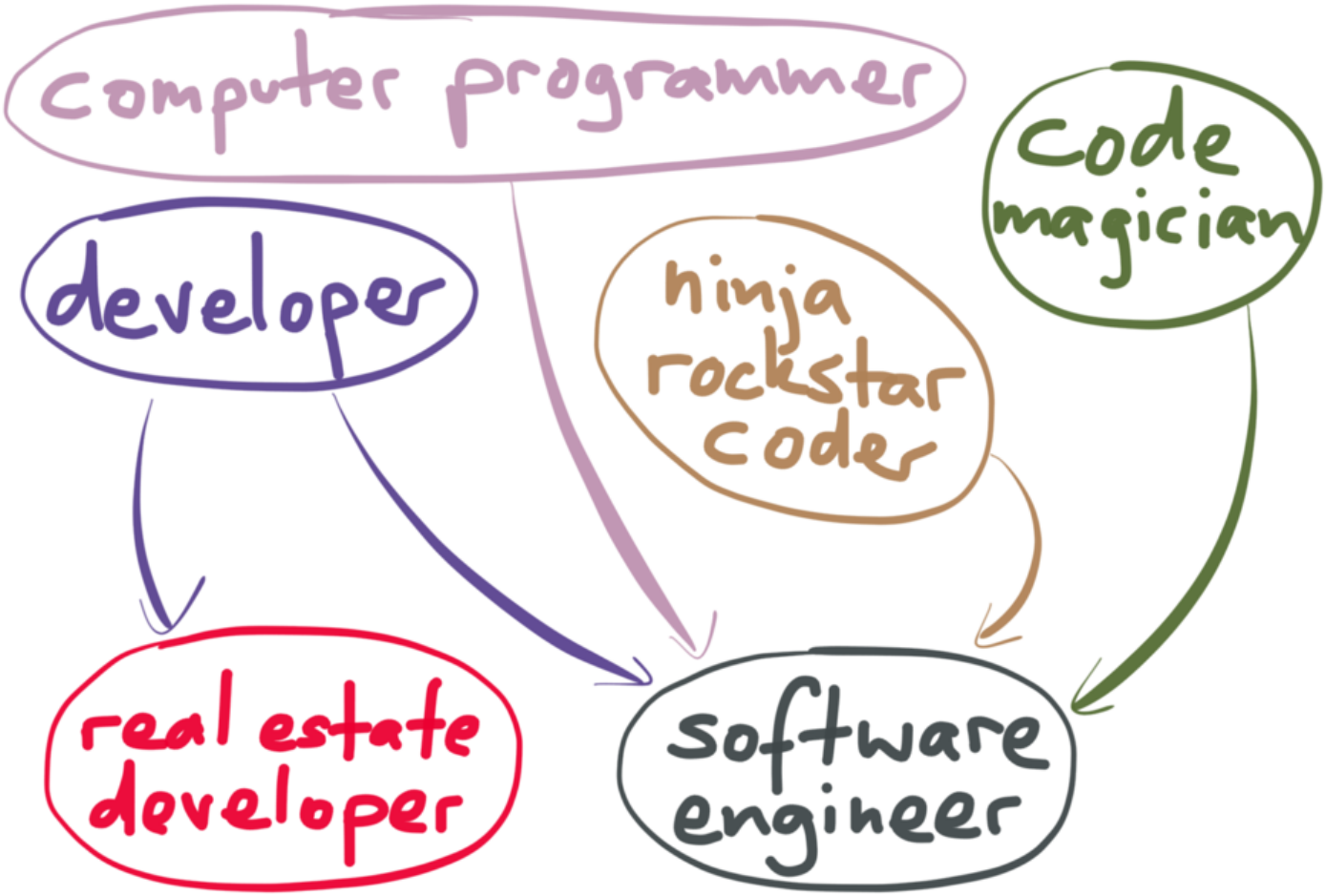


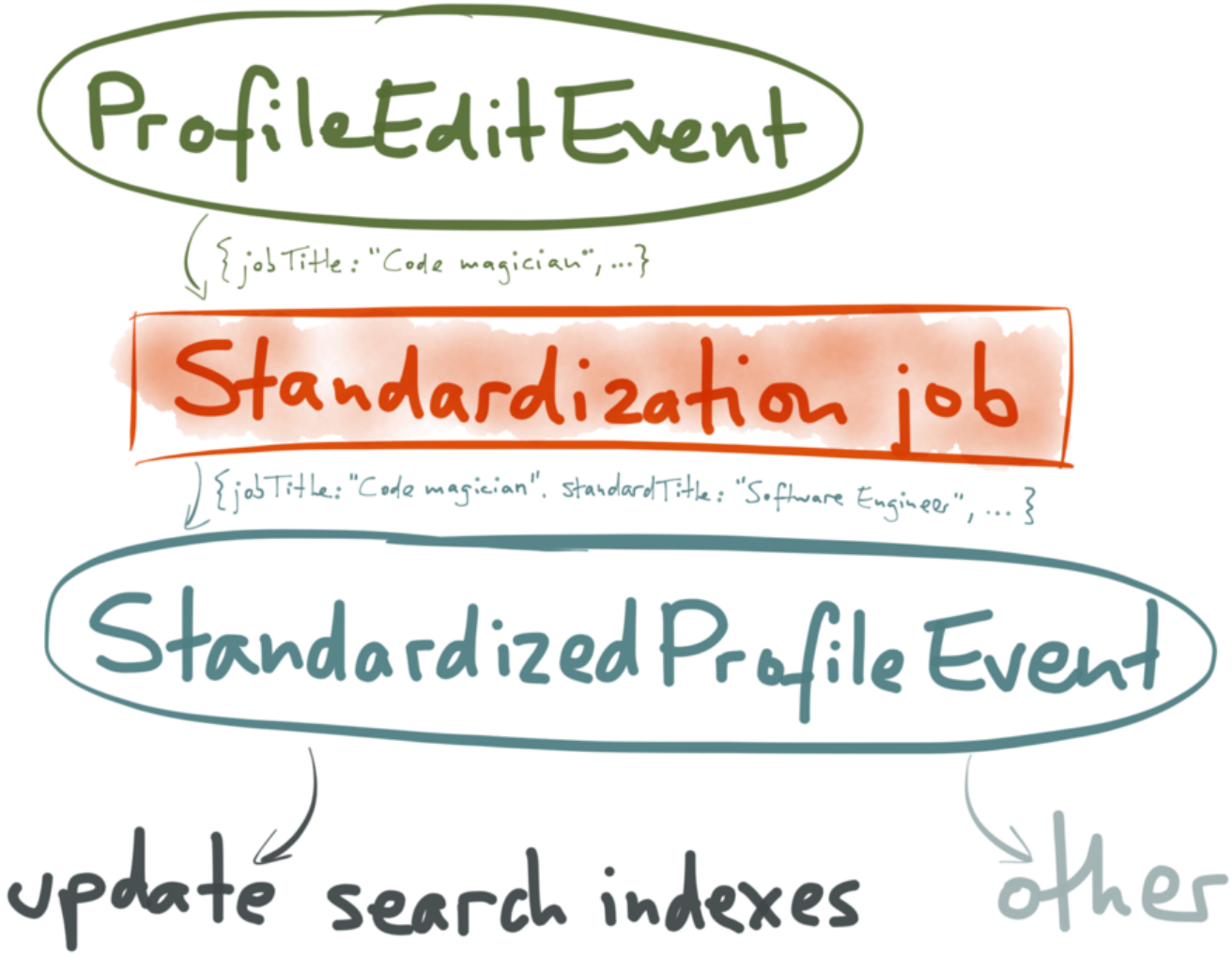
2. Data standardization

(s13n?)


People
Jobs
Keywords
First Name
Last Name
Title
Company
School
Location
Anywhere
Search
Reset
Advanced People Search
Reset
Close
Relationship
1st Connections
2nd Connections
Group Members
3rd + Everyone Else
Location
Cambridge, United Kingdom
+ Add
Current Company
Industry
Past Company
School
Profile Language
Nonprofit Interests
Groups
Years of Experience
Function
Academics
Accounting
Administrative
Business development
Buyer
Seniority Level
Manager
Owner
Partner
CXO
VP
Interested In
Company Size
Fortune
Fortune 50
Fortune 51-100
Fortune 101-250
Fortune 251-500
Fortune 501-1000
When Joined


3rd + Everyone Else
Location
Cambridge, United Kingdom
+ Add
Current Company
Industry
Past Company
School
Profile Language
Nonprofit Interests
Function
Academics
Accounting
Administrative
Business development
Buyer
Seniority Level
Manager
Owner
Partner
CXO
VP
Interested In
Company Size
Fortune
Fortune 50
Fortune 51-100


computer programmer
developer
ninja
rockstar
coder
code
magician
real estate
developer
software
engineer


ProfileEditEvent
{jobTitle: "Code magician", ...}
Standardization job
{jobTitle: "Code magician", standardTitle: "Software Engineer", ...}
StandardizedProfileEvent
update search indexes
other


ProfileEditEvent
Magic
(offline learning/analysis)
{jobTitle: "Code magician", ...}
Standardization job
{jobTitle: "Code magician", standardTitle: "Software Engineer", ...}
StandardizedProfileEvent
update search indexes
other

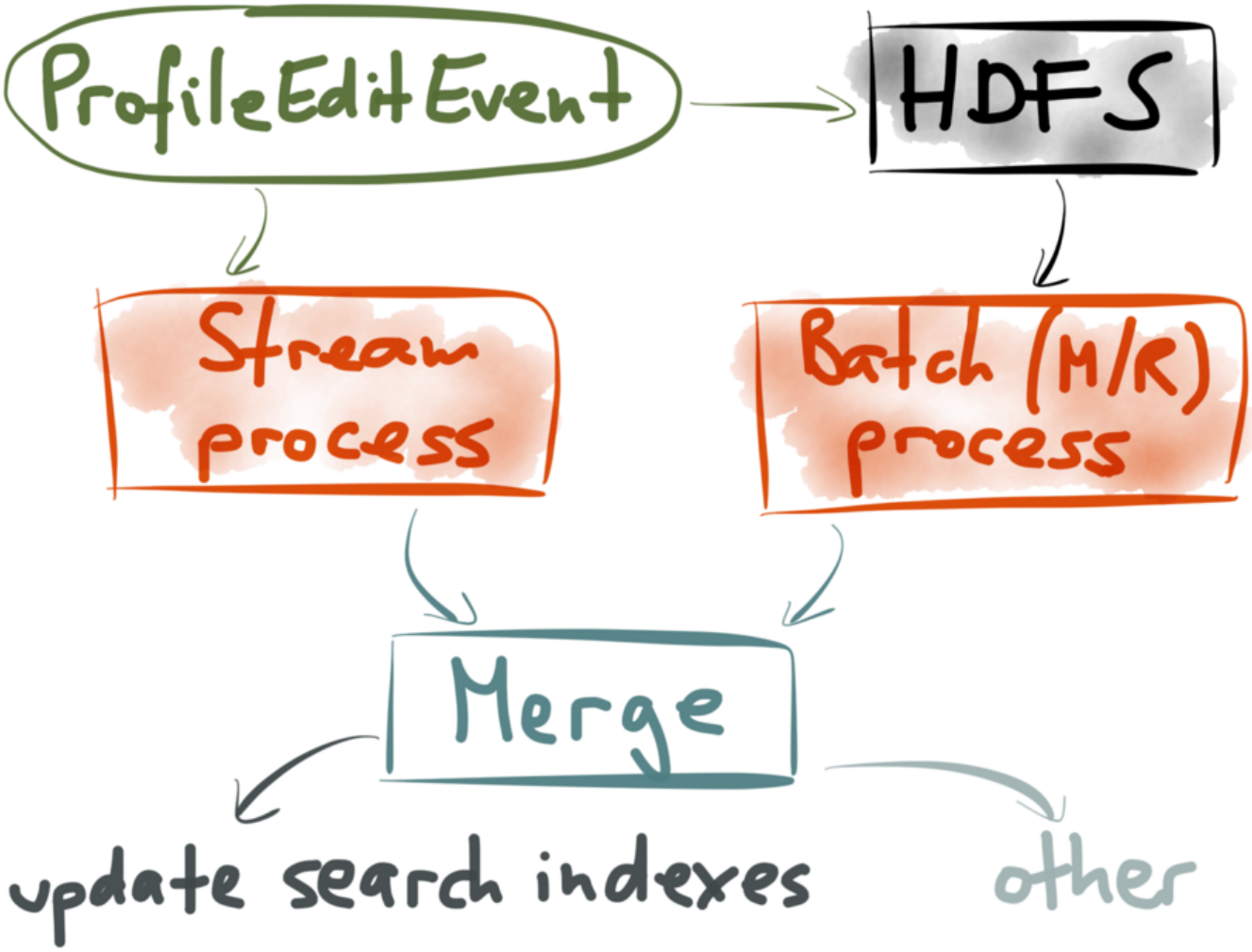
Near real-time index updates

⇒ Stream process

Standardization rules change

⇒ Re-classify everything

⇒ Batch process

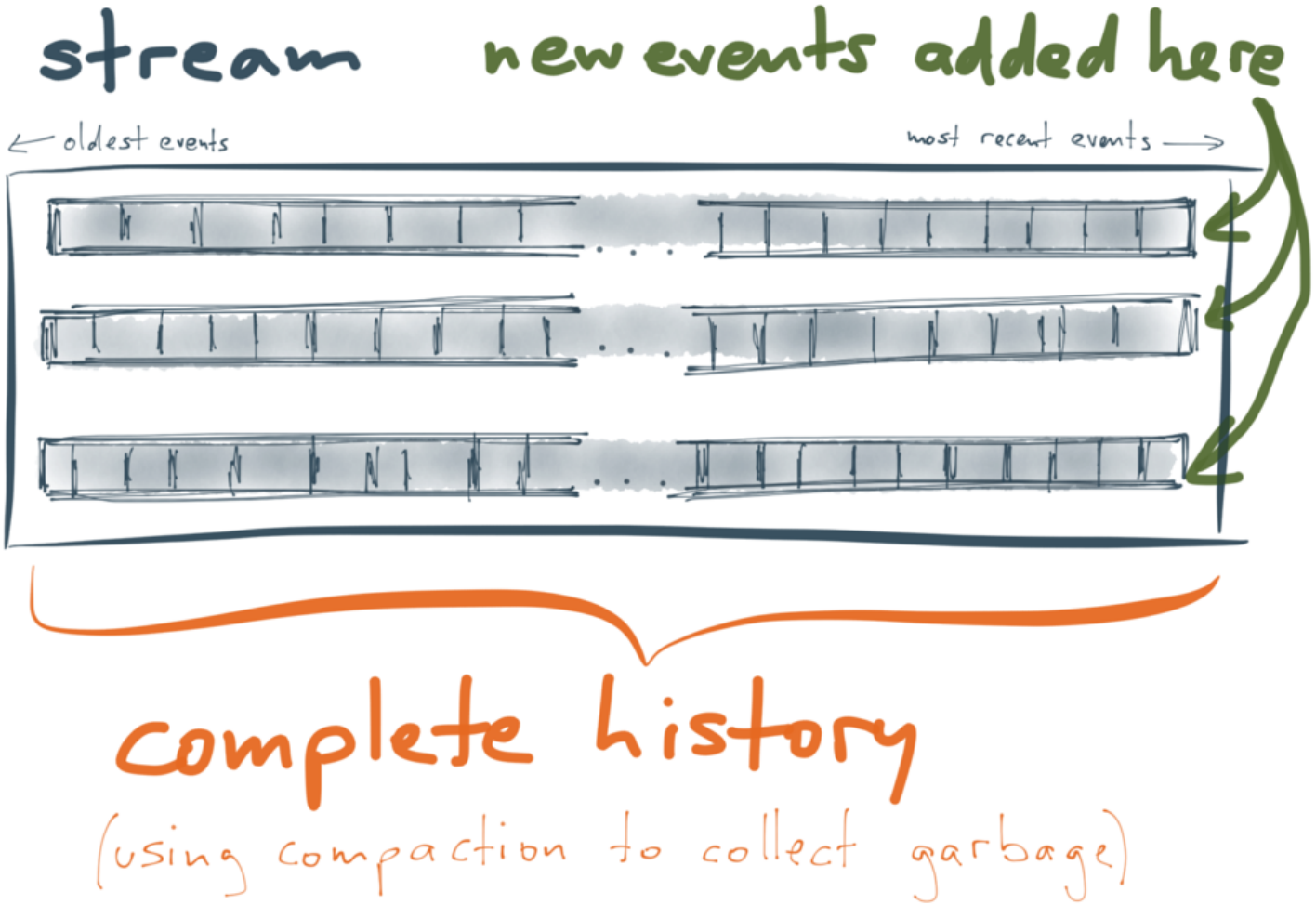

ProfileEditEvent
HDFS
Stream process
Batch (M/R) process
Merge
update search indexes
other

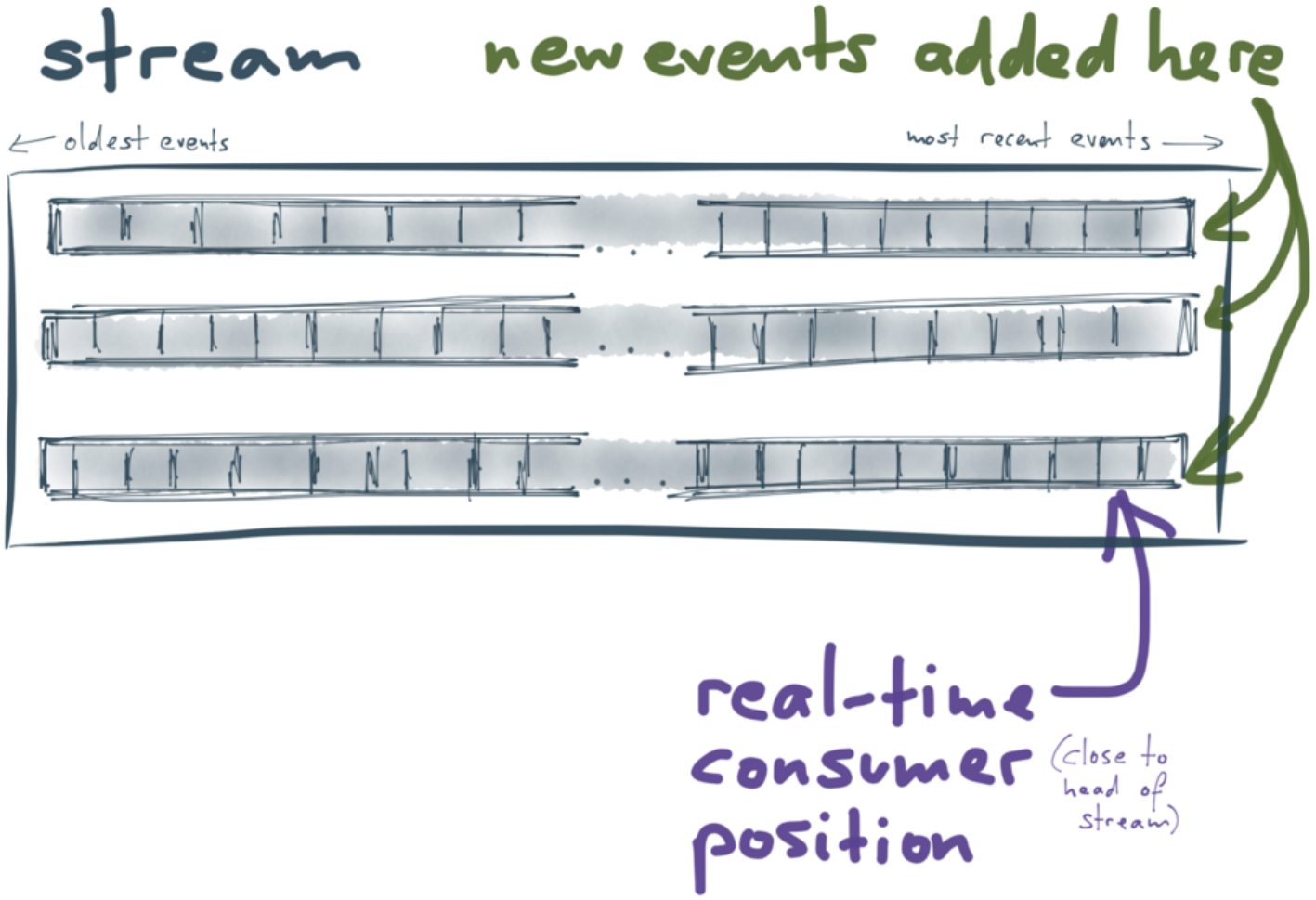
λ architecture

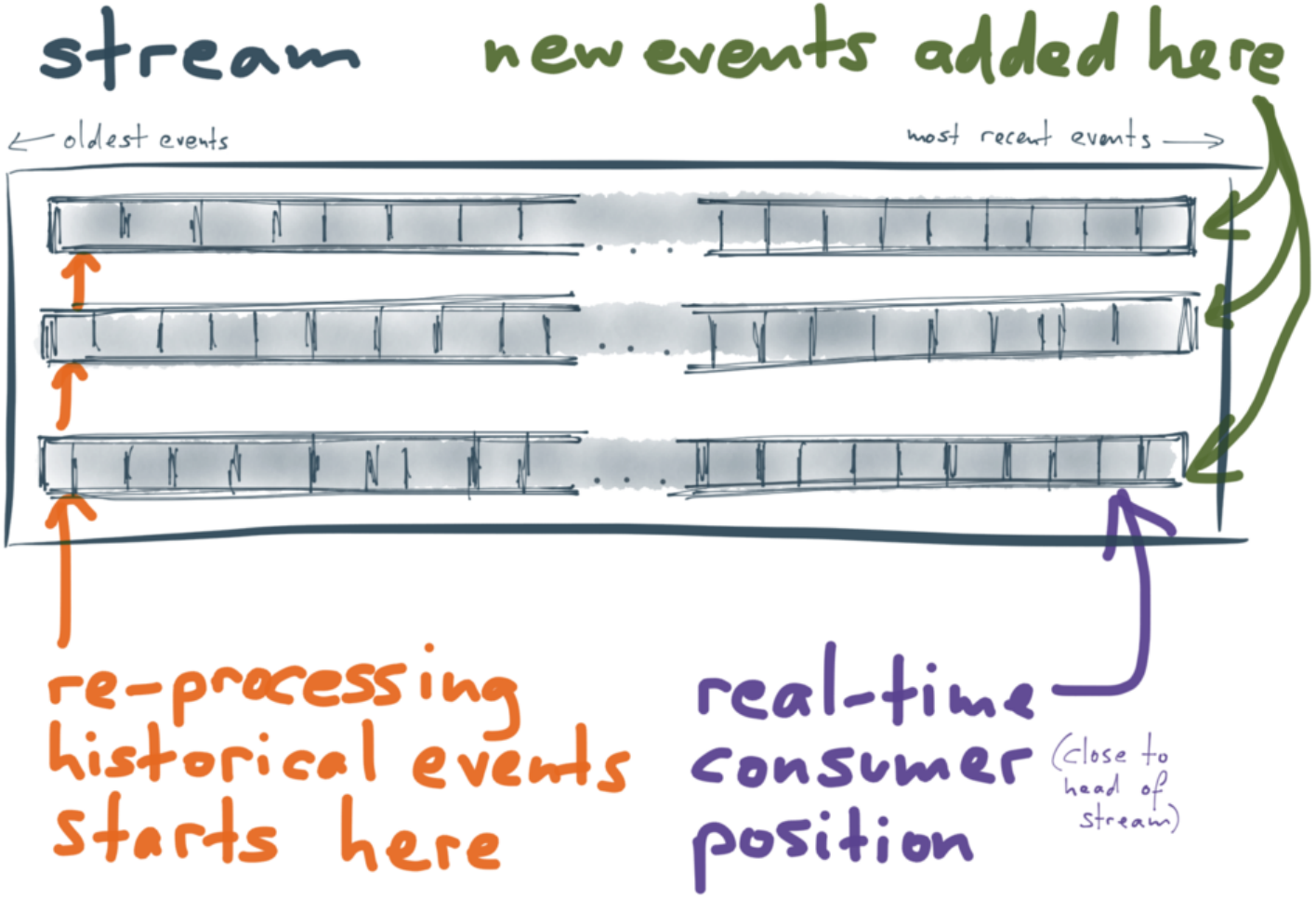
"implement the same job twice"

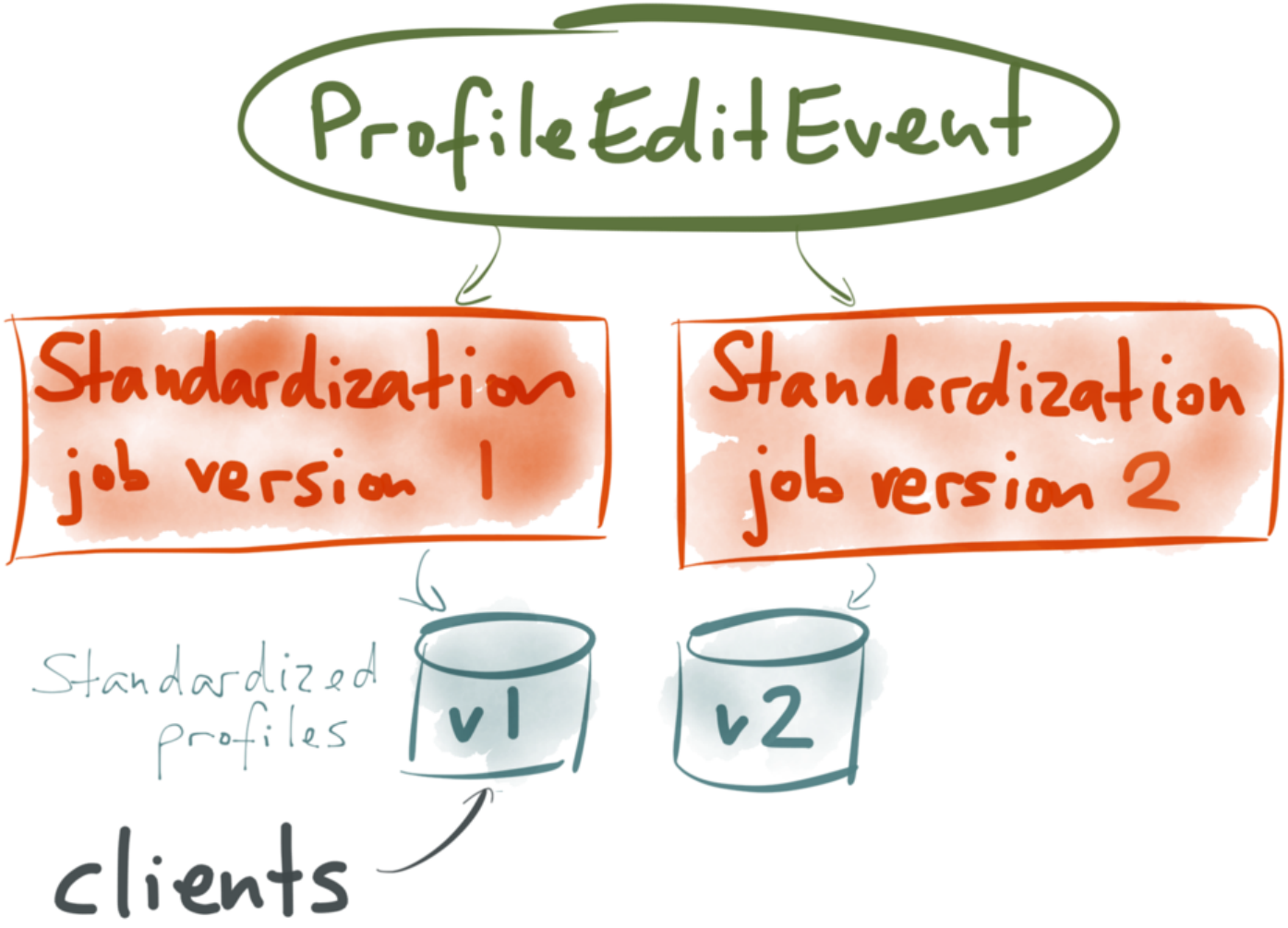
K architecture

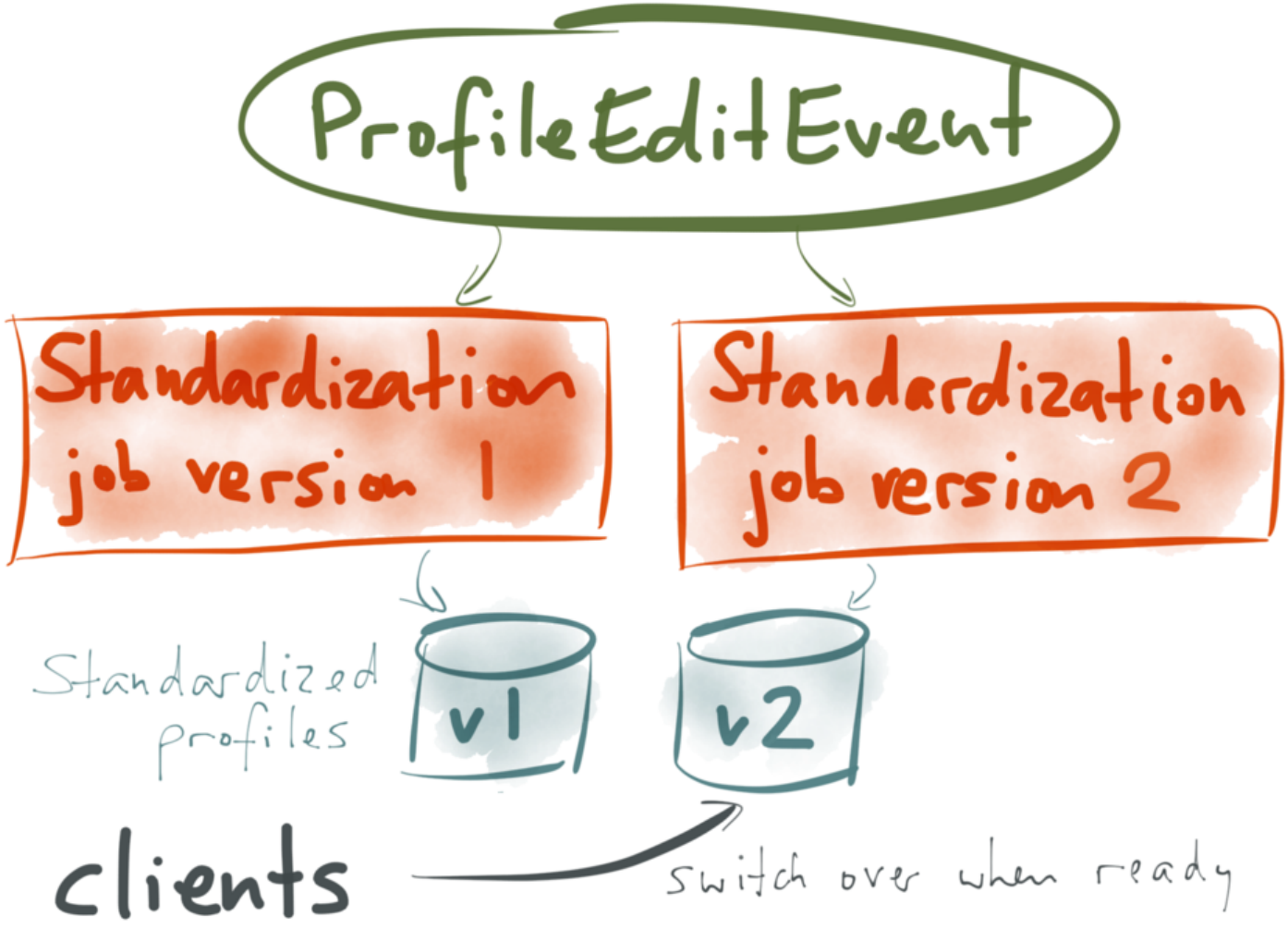
"process real-time data and re-process historical data in the same framework"


stream
new events added here
← oldest events
most recent events →
complete history
(using compaction to collect garbage)


stream
new events added here
← oldest events
most recent events →
real-time consumer position
(close to head of stream)


stream
new events added here
← oldest events
most recent events →
re-processing historical events starts here
real-time consumer position
(close to head of stream)


ProfileEditEvent
Standardization job version 1
Standardization job version 2
Standardized profiles
v1
v2
clients


ProfileEditEvent
Standardization job version 1
Standardization job version 2
Standardized profiles
v1
v2
clients
switch over when ready

~Recap~

JOIN streams to replicated table

RE-PROCESS streams when processing logic changes

High throughput & good operability

References (fun stuff to read)

1. Martin Kleppmann: "Designing data-intensive applications." O'Reilly Media, to appear in 2015. http://dataintensive.net
2. Jay Kreps: "Why local state is a fundamental primitive in stream processing." 31 July 2014. http://radar.oreilly.com/2014/07/why-local-state-is-a-fundamental-primitive-in-stream-processing.html
3. Jay Kreps: "I ♥ Logs." O'Reilly Media, September 2014. http://shop.oreilly.com/product/0636920034339.do
4. Nathan Marz and James Warren: "Big Data: Principles and best practices of scalable realtime data systems." Manning MEAP, to appear January 2015. http://manning.com/marz/
5. Jakob Homan: "Real time insights into LinkedIn's performance using Apache Samza." 18 Aug 2014. http://engineering.linkedin.com/samza/real-time-insights-linkedins-performance-using-apache-samza
6. Martin Kleppmann: "Moving faster with data streams: The rise of Samza at LinkedIn." 14 July 2014. http://engineering.linkedin.com/stream-processing/moving-faster-data-streams-rise-samza-linkedin
7. Praveen Neppalli Naga: "Real-time Analytics at Massive Scale with Pinot." 29 Sept 2014. http://engineering.linkedin.com/analytics/real-time-analytics-massive-scale-pinot
8. David He: "Monitor and Improve Web Performance Using RUM Data Visualization." 19 Sept 2014. http://engineering.linkedin.com/performance/monitor-and-improve-web-performance-using-rum-data-visualization
9. Lili Wu, Sam Shah, Sean Choi, Mitul Tiwari, and Christian Posse: "The Browsemaps: Collaborative Filtering at LinkedIn," at 6th Workshop on Recommender Systems and the Social Web, Oct 2014. http://ls13-www.cs.uni-dortmund.de/homepage/rsweb2014/papers/rsweb2014_submission_3.pdf
10. Shirshanka Das, Chavdar Botev, Kapil Surlaker, et al.: "All Aboard the Databus!," at ACM Symposium on Cloud Computing (SoCC), October 2012. http://www.socc2012.org/s18-das.pdf
11. Apache Samza documentation. http://samza.incubator.apache.org