

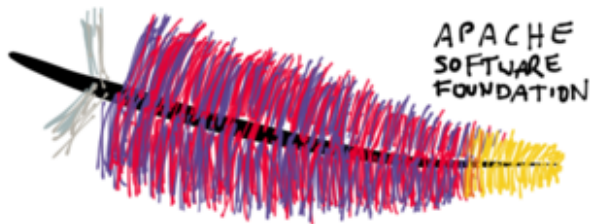
SCALABLE **REAL-TIME** DATA PROCESSING WITH



Kafka

&

Samza



Martin Kleppmann
@martinkl

dataintensive.net



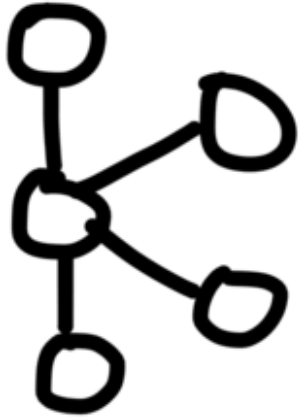
@martinkl

samza.apache.org

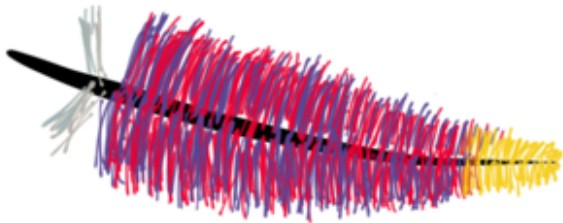
samza

DISTRIBUTED STREAM PROCESSING





Kafka

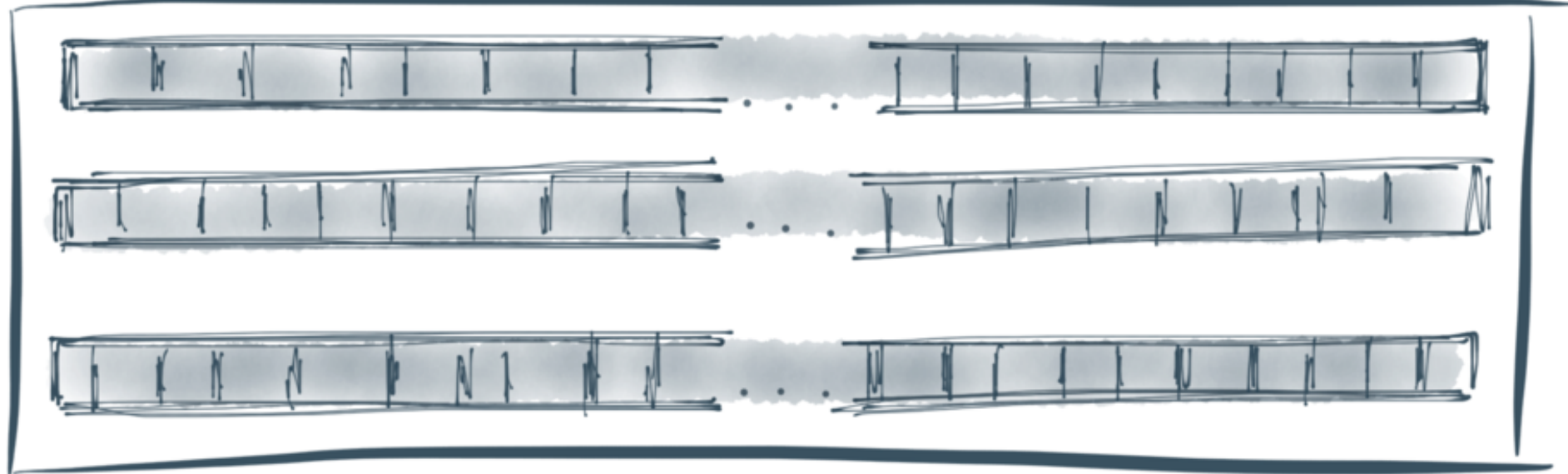


APACHE
SOFTWARE
FOUNDATION

stream

← oldest events

most recent events →

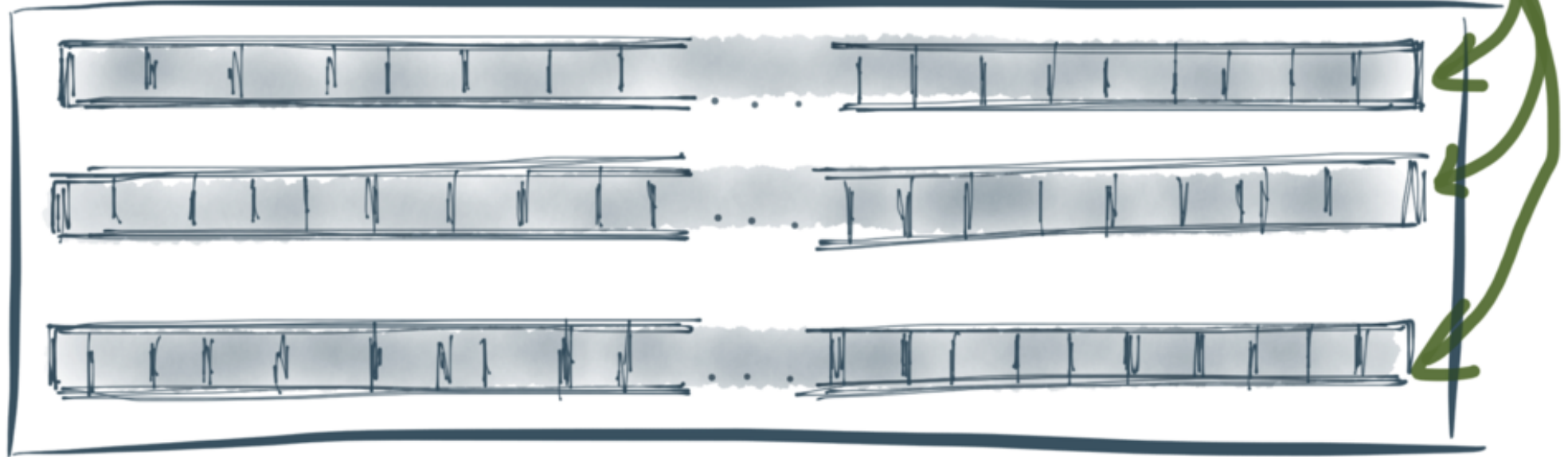


stream

new events added here

← oldest events

most recent events →

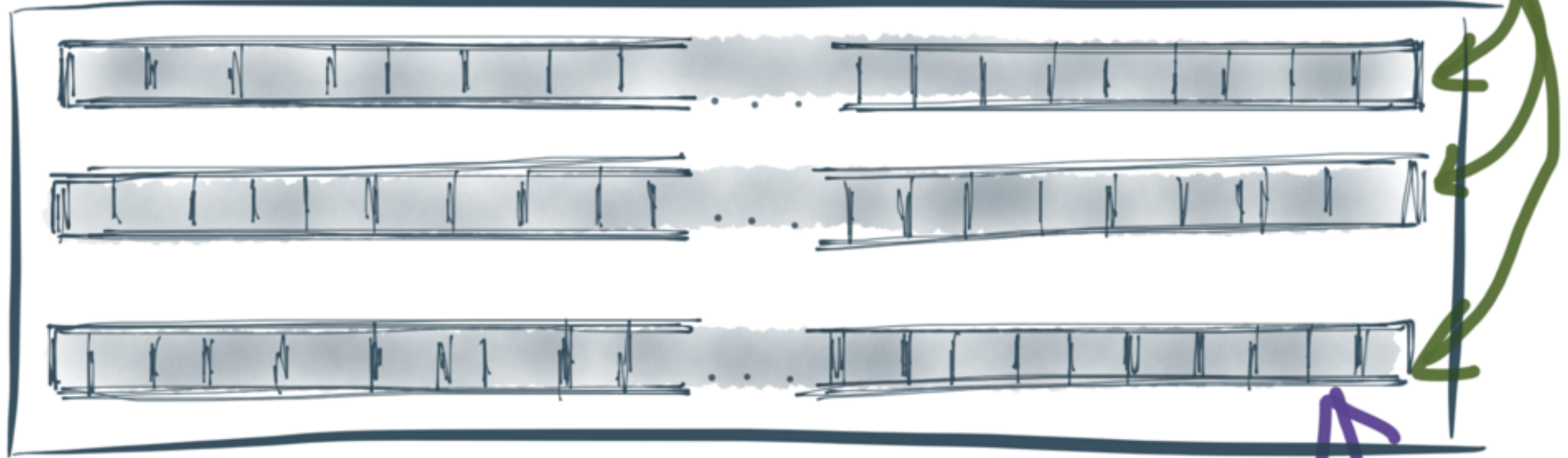


stream

new events added here

← oldest events

most recent events →



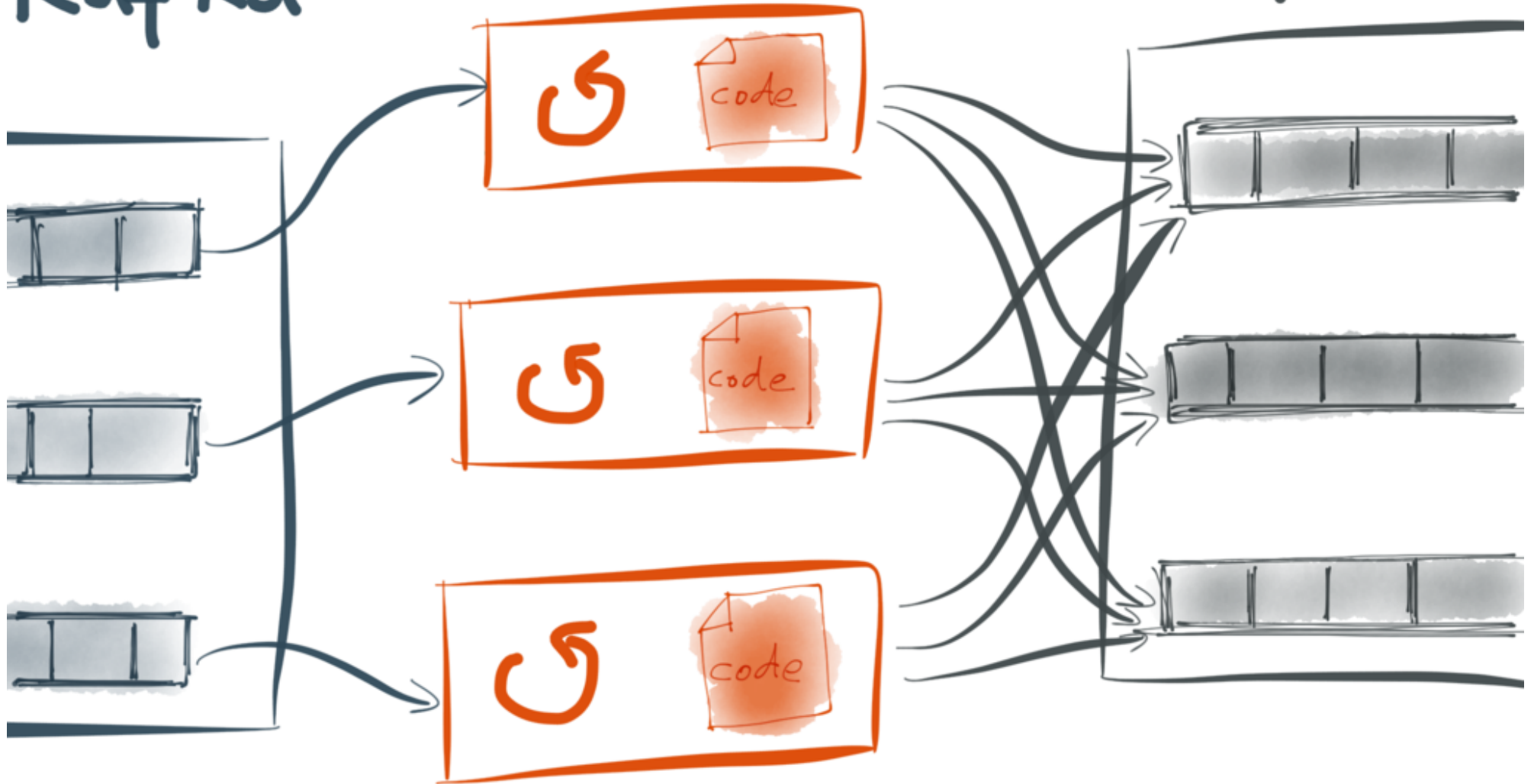
consumer
position

(close to
head of
stream)

Kafka

Samza

Kafka



```
public interface StreamTask {  
    void process(  
        IncomingMessageEnvelope msg,  
        MessageCollector collector,  
        TaskCoordinator coordinator  
    );  
}
```



YARN

Samza

tez

MR

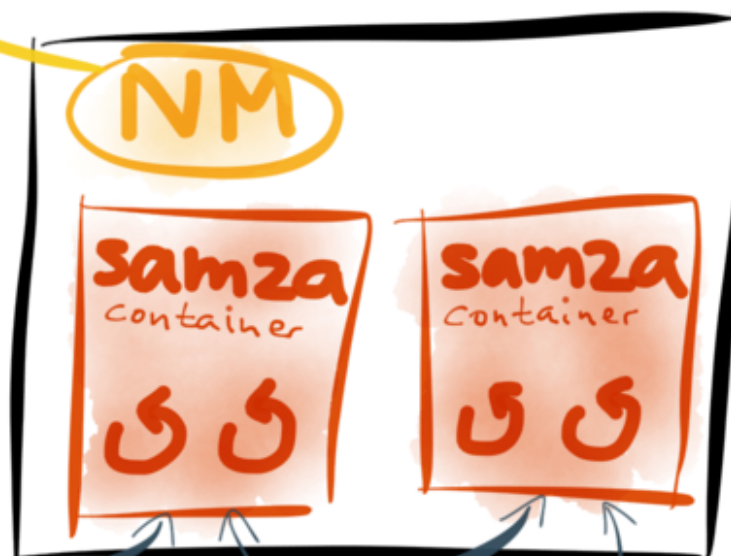
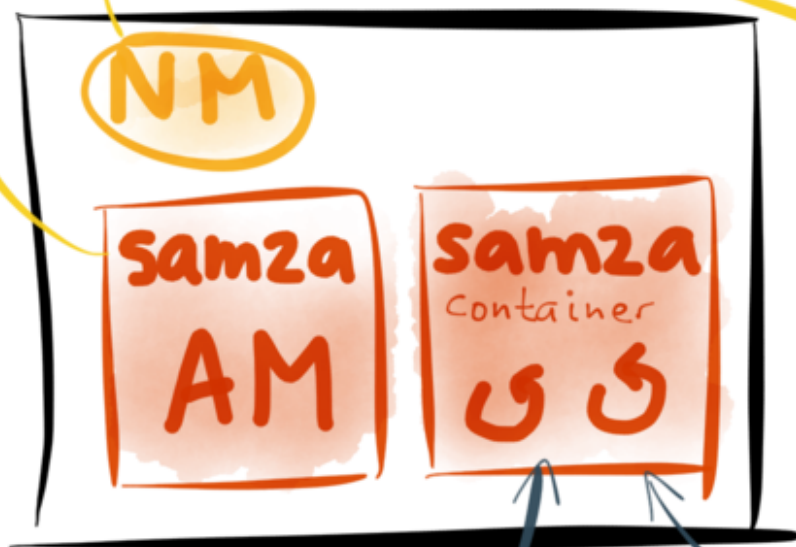


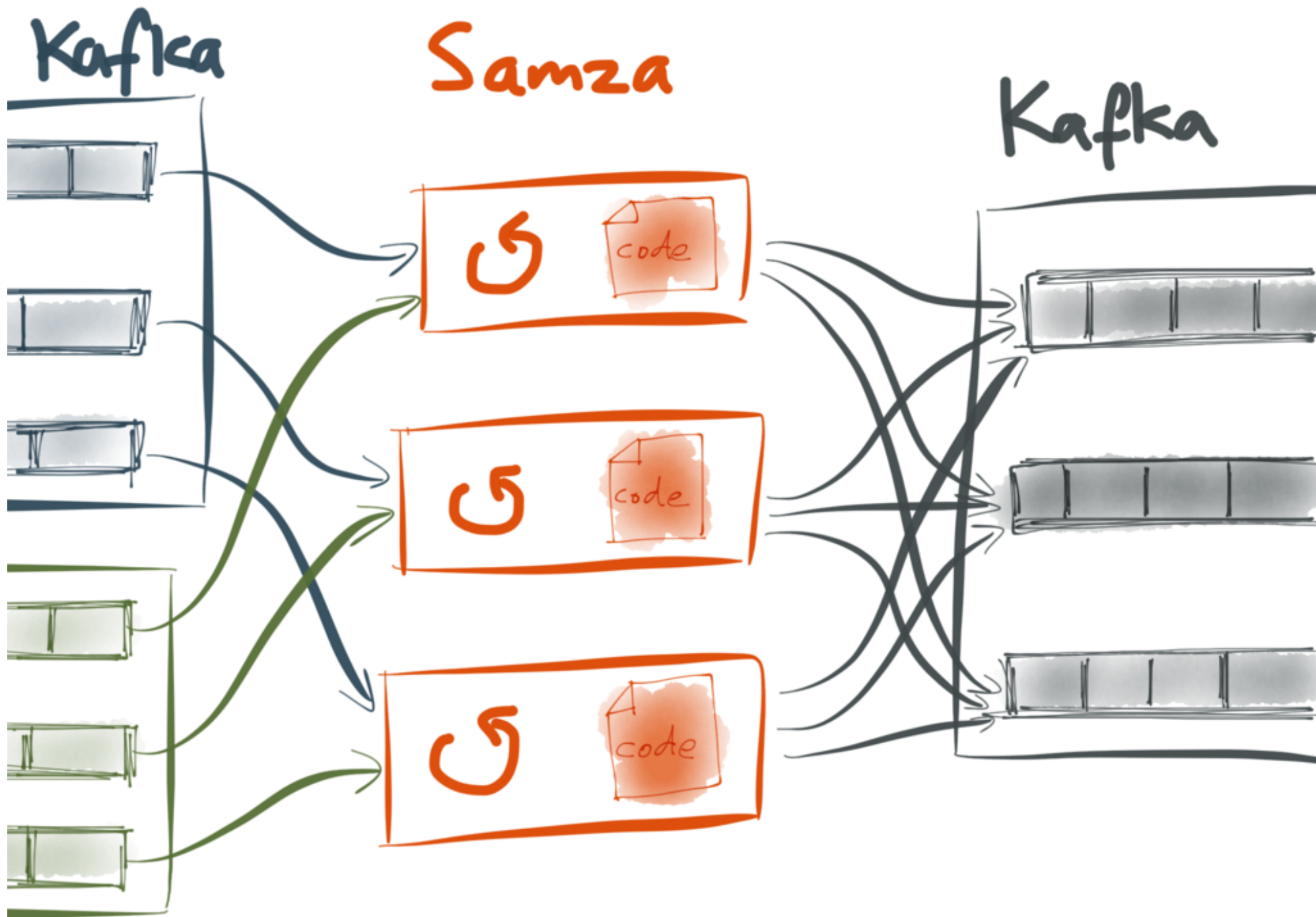
Samza

etc...

HDFS

Kafka





PageViewEvent

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

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 ^

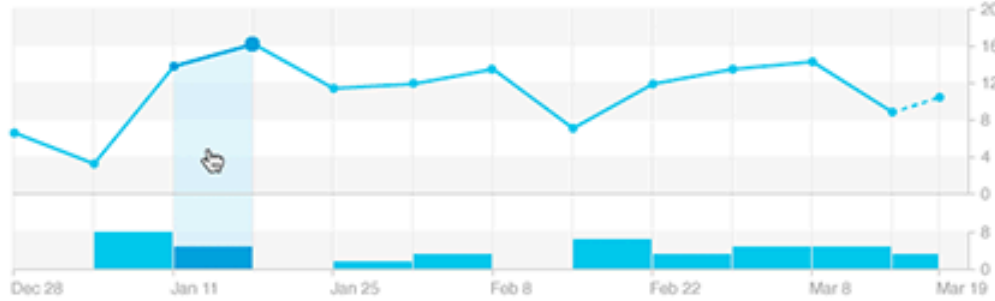
93 Profile views
Last 90 days

13 Viewers from the San Francisco Bay Area

7 Viewers from the Design Industry

9 Viewers found you from Google

Profile views Actions taken



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

```
graph TD; A([PageViewEvent]) --> B[People also viewed]; A --> C[Who viewed your profile]; A --> D[A/B tests]; A --> E[Reporting]; A --> F[Site speed monitoring]; A --> G[Relevance models];
```

People
also viewed

Who viewed
your profile

A/B
tests

Reporting

Site speed
monitoring

Relevance
models

PageViewEvent

```
graph TD; A([PageViewEvent]) --> B[Who viewed your profile]; A --> C[People also viewed]; A --> D[A/B tests]; A --> E[Relevance models]; A --> F[Site speed monitoring]; A --> G[Reporting];
```

People
also viewed

Relevance
models

Site speed
monitoring

Reporting

A/B
tests

Who viewed
your profile

Who's viewed your profile

How you rank for profile views

Hide charts ^

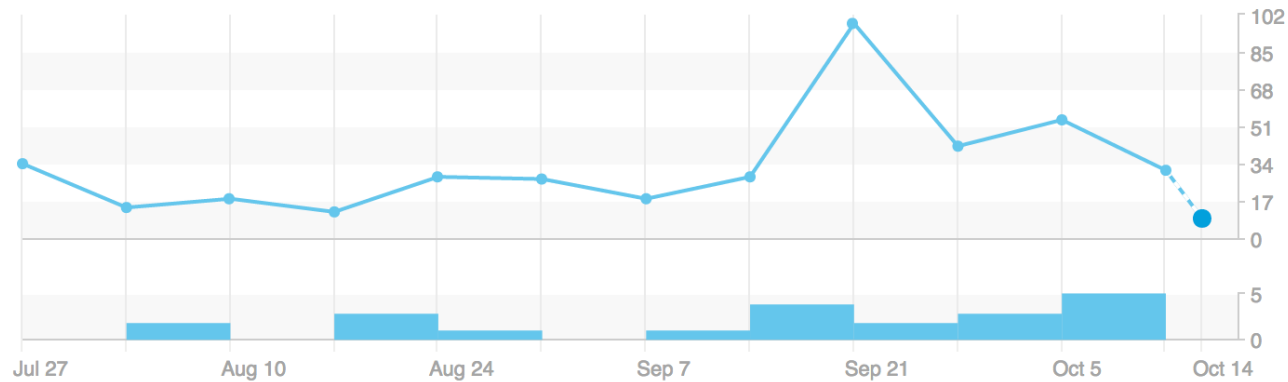
411 Profile views
Last 90 days

98 Viewers found you from
People similar to you

27 Viewers who work at LinkedIn

70 Viewers with the title Software
Developer

— Profile Views Actions Taken



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](#)

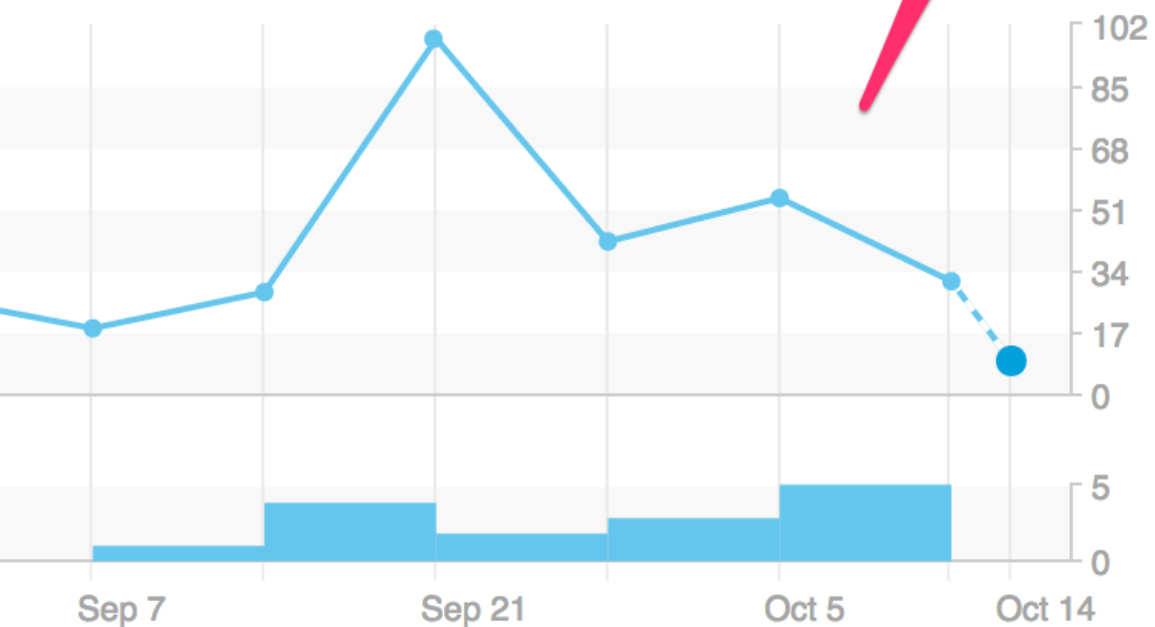
How you rank for profile views

Hide charts ^

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](#) ▶

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

{viewerId: 1234}

(enrich with profile data

{viewerId: 1234, viewerProfile: {
jobTitle: "Software Engineer"}}

`{viewerId: 1234}`

(enrich with profile data

`{viewerId: 1234, viewerProfile: {
jobTitle: "Software Engineer"}}`

indexing

70

Viewers with the title Software
Developer

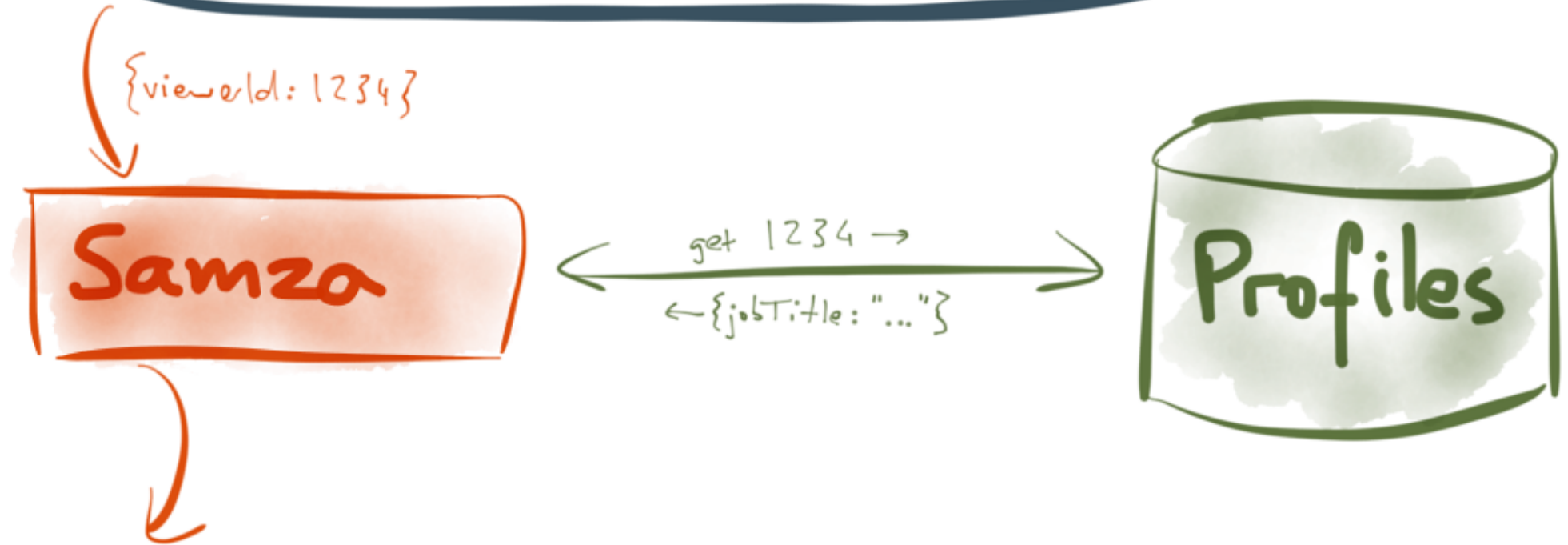
PageViewEvent stream

{viewId: 1234}

Samza

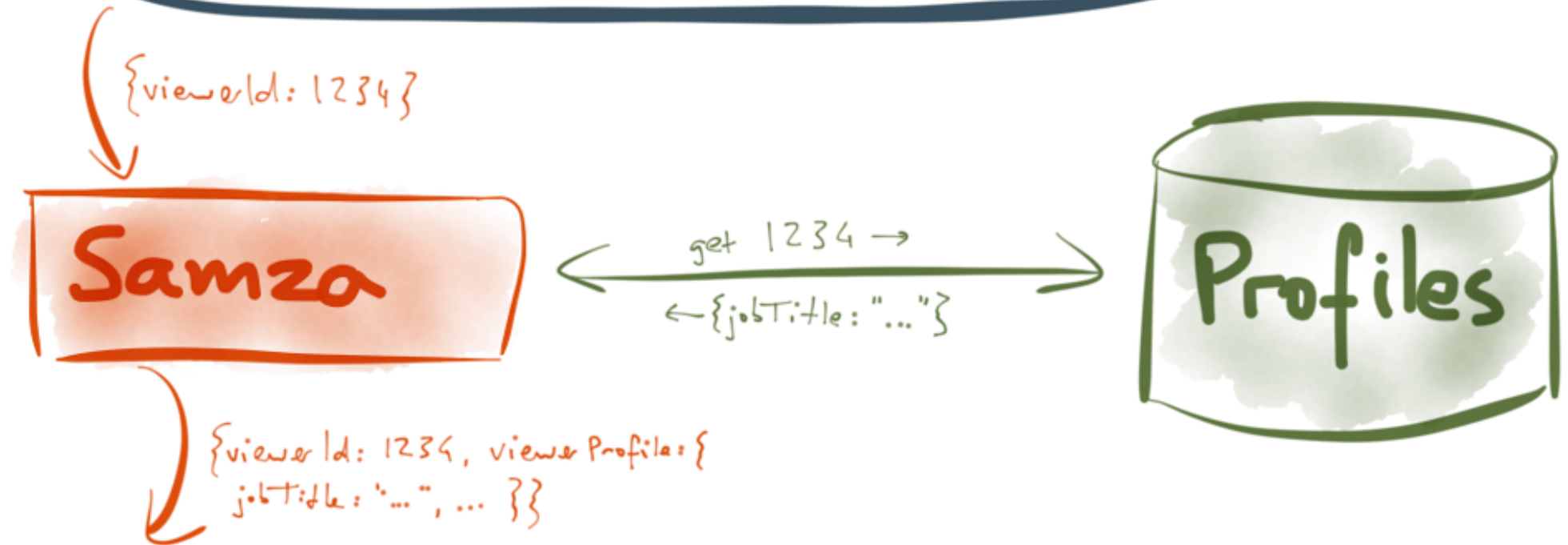
PageViewEventWithViewerProfile

PageViewEvent stream



PageViewEventWithViewerProfile

PageViewEvent stream



PageViewEventWithViewerProfile

PageViewEvent stream

~100-500k msg/sec

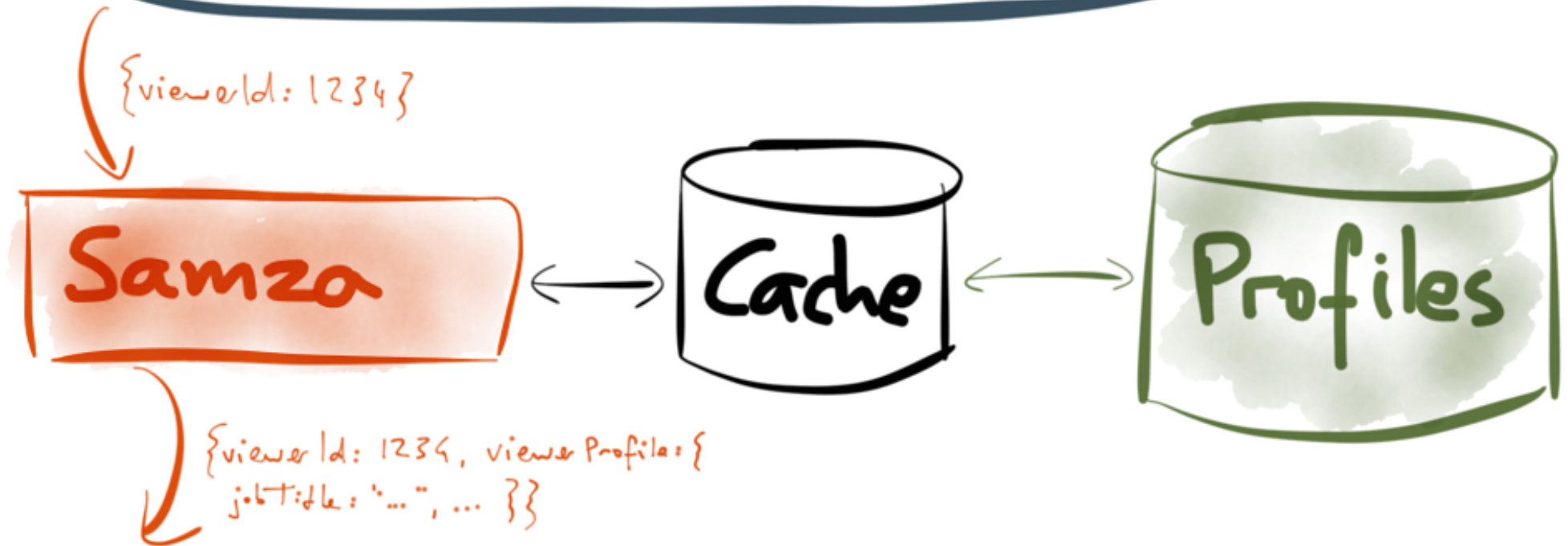
Samza

~1-5k req/sec

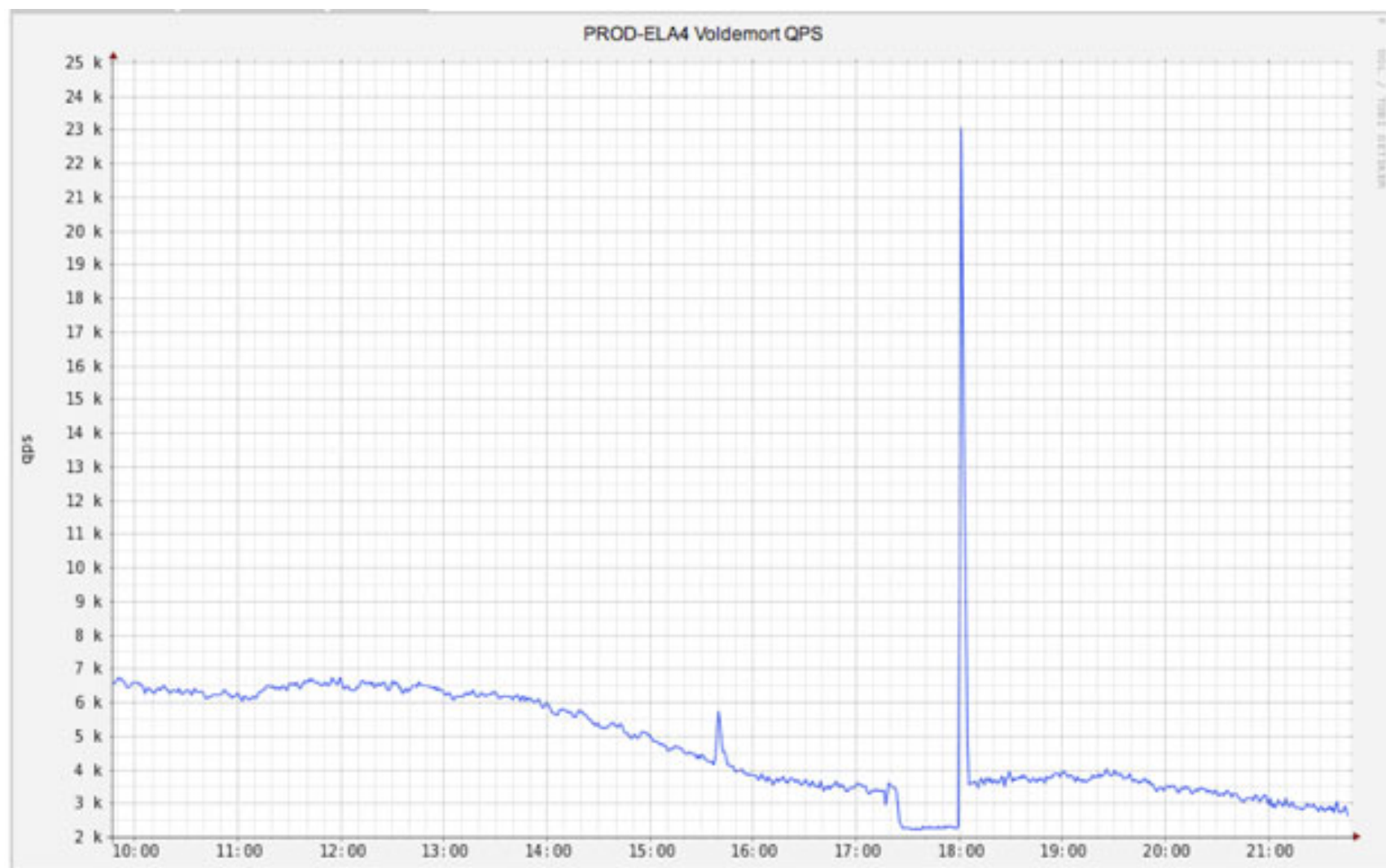
Profiles

PageViewEventWithViewerProfile

PageViewEvent stream



PageViewEventWithViewerProfile



key = **urn:linkedin:profile:1234**

value = {

 eventType: **ProfileEditEvent**,

 timestamp: 1413215518,

 profile: {

 location: "Cambridge, UK",

 industry: "Software",

 positions: [

 {job_title: "Author", company: "O'Reilly"},

 ...

]}}

ProfileEditEvent

Samza

ProfileEditEvent

$\{profileId: 1234, old: \{...\}, new: \{...\}\}$

Samza

set 1234, $\{...\}$

Profiles
(a copy just for us!)

PageViewEvent

ProfileEditEvent

{profileId: 1234, old: {...}, new: {...}}

Samza

set 1234, {...}

Profiles
(a copy just for us!)

PageViewEvent

{viewerId:1234}

Samza

ProfileEditEvent

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

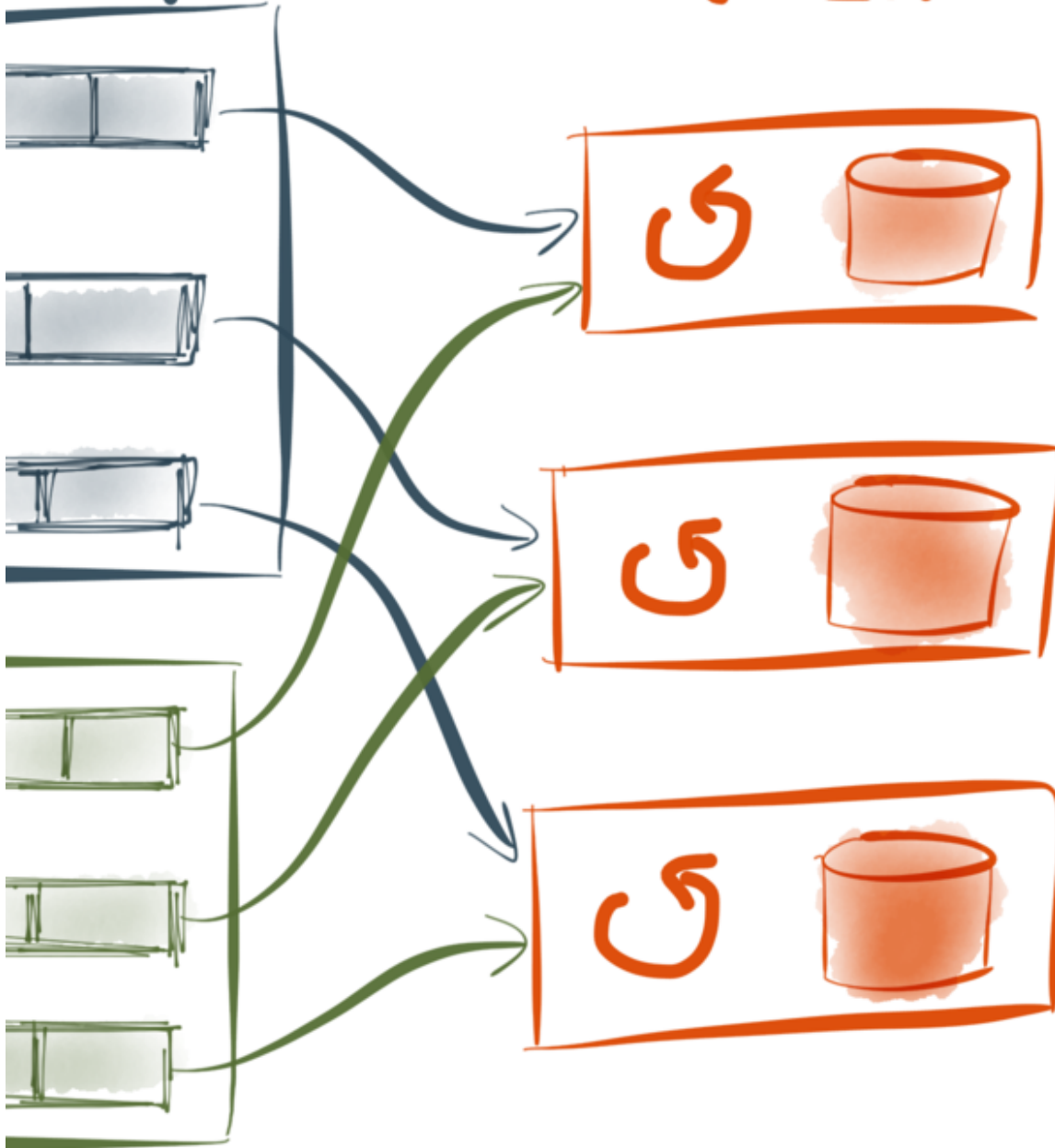
ProfileEditEvent



PageViewEventWithViewerProfile

Kafka

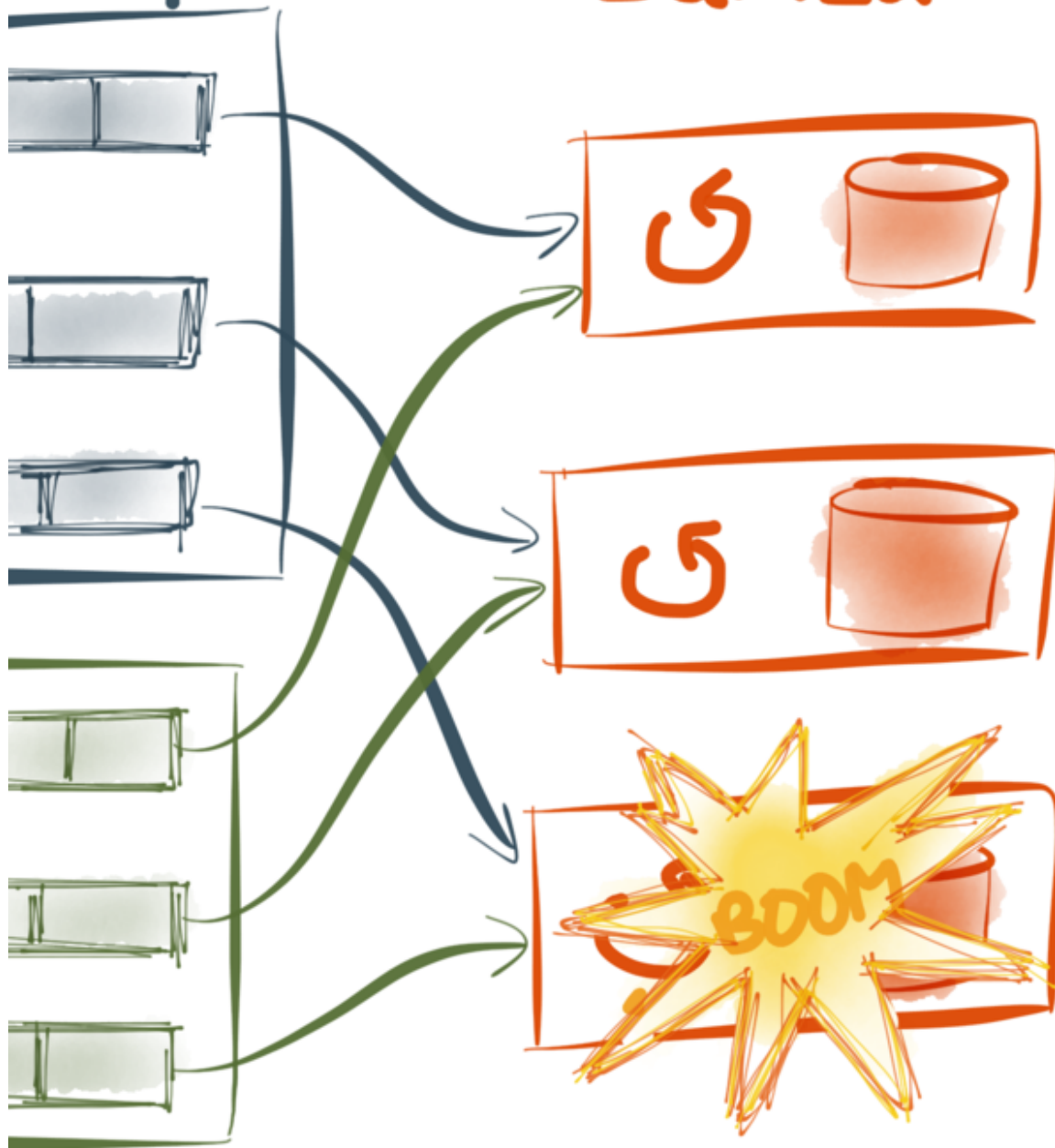
Samza



Embedded
database
(Level DB/
Rocks DB)

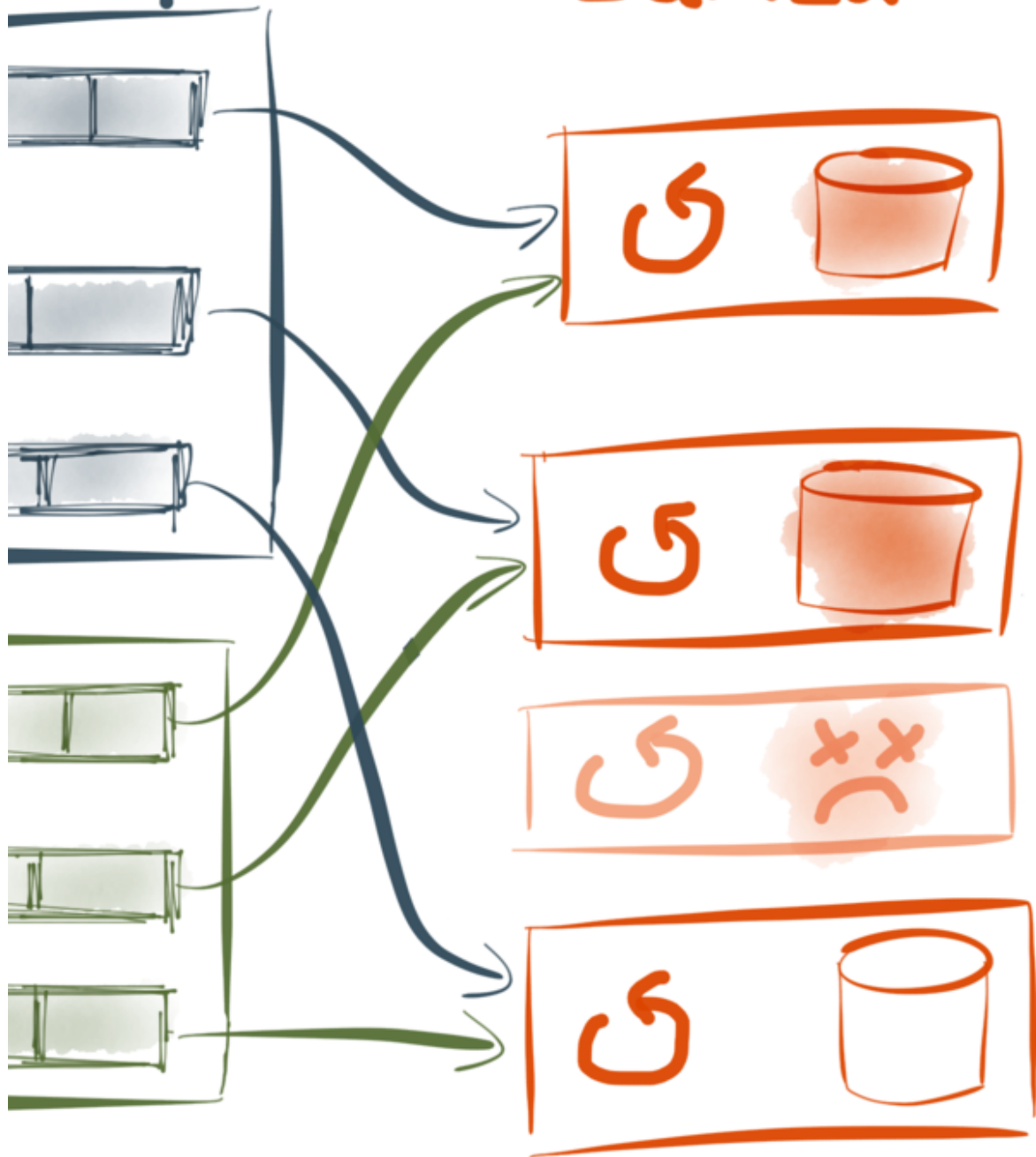
Kafka

Samza



Kafka

Samza



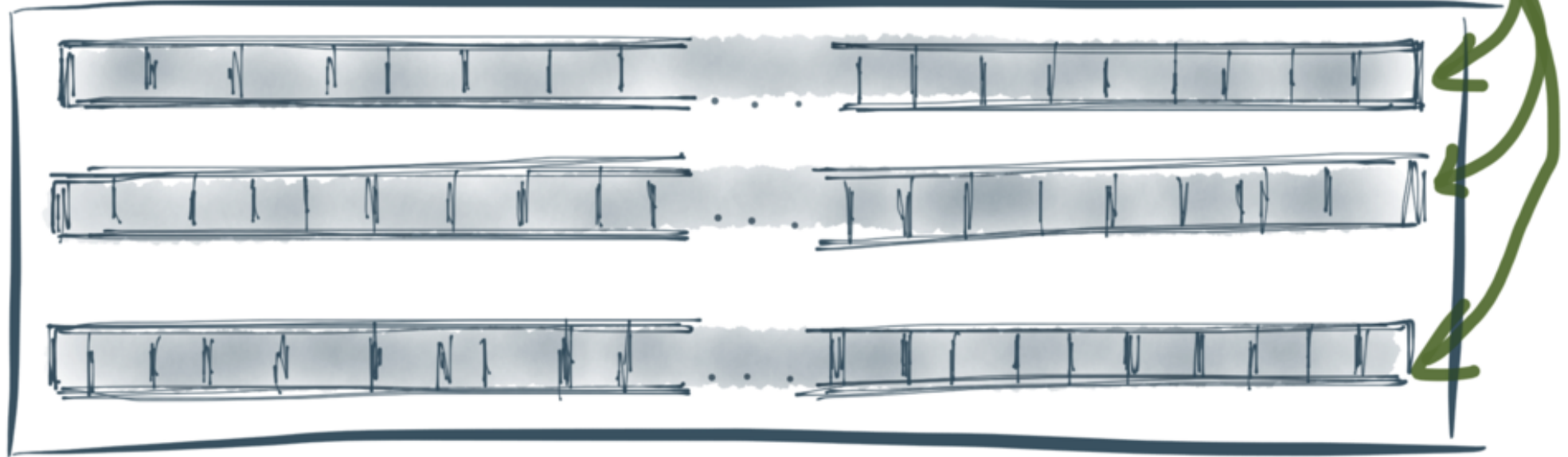
Data lost 😞

stream

new events added here

← oldest events

most recent events →



complete history

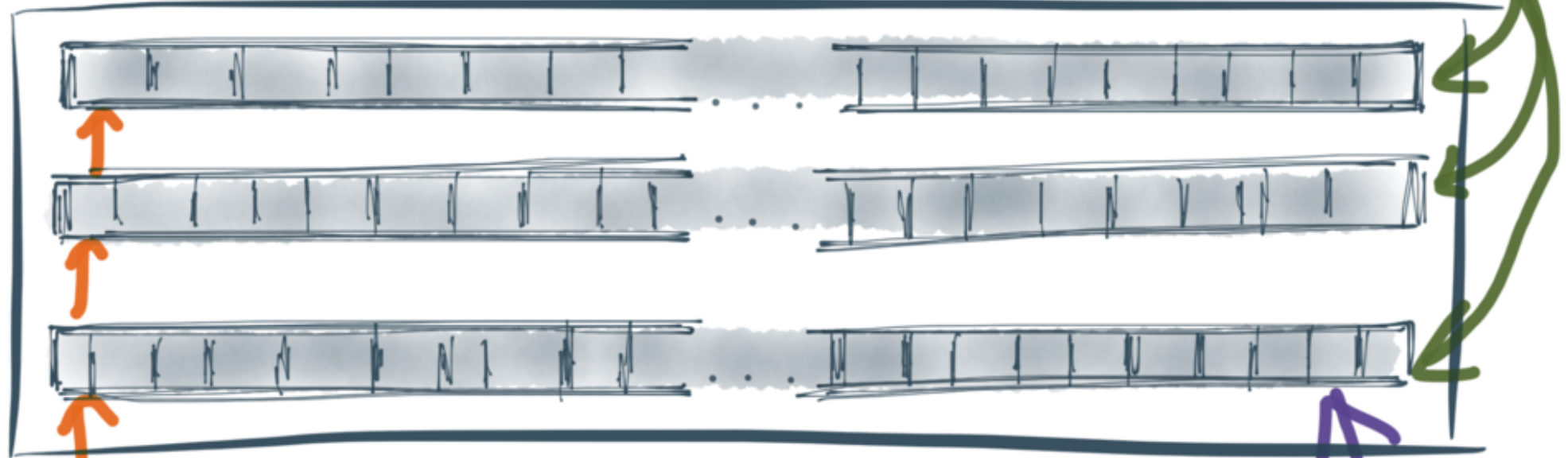
(using compaction to collect garbage)

stream

new events added here

← oldest events

most recent events →



re-processing
historical events
starts here

consumer
position

(close to
head of
stream)

A	42
---	----

B	21
---	----

A	44
---	----

C	12
---	----

A	45
---	----

Kafka changelog compaction



key = **urn:linkedin:profile:1234**

value = {

 eventType: **ProfileEditEvent**,

 timestamp: 1413215518,

 profile: {

 location: “Cambridge, UK”,

 industry: “Software”,

 positions: [

 {job_title: “Author”, company: “O’Reilly”},

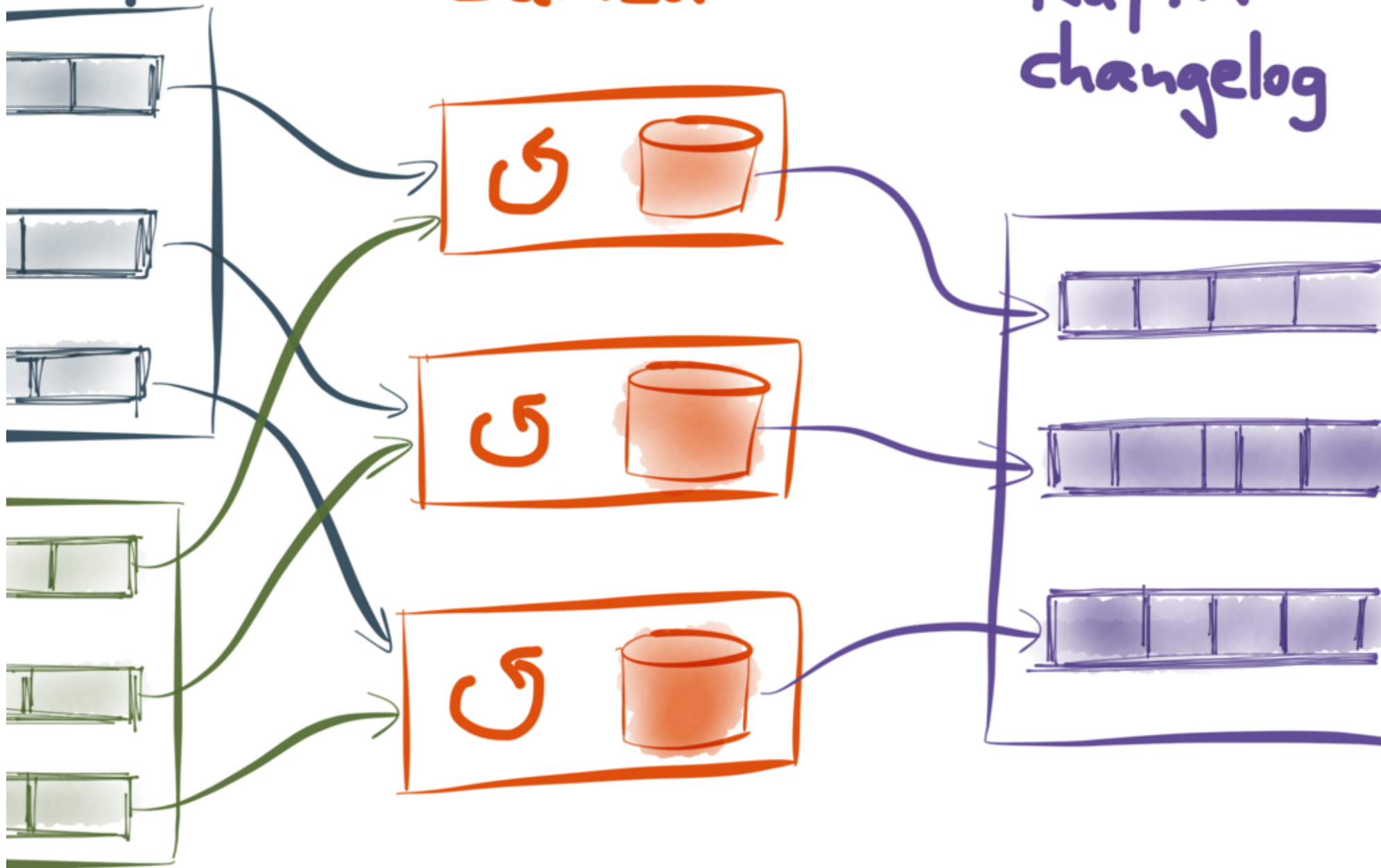
 ...

]}}

Kafka

Samza

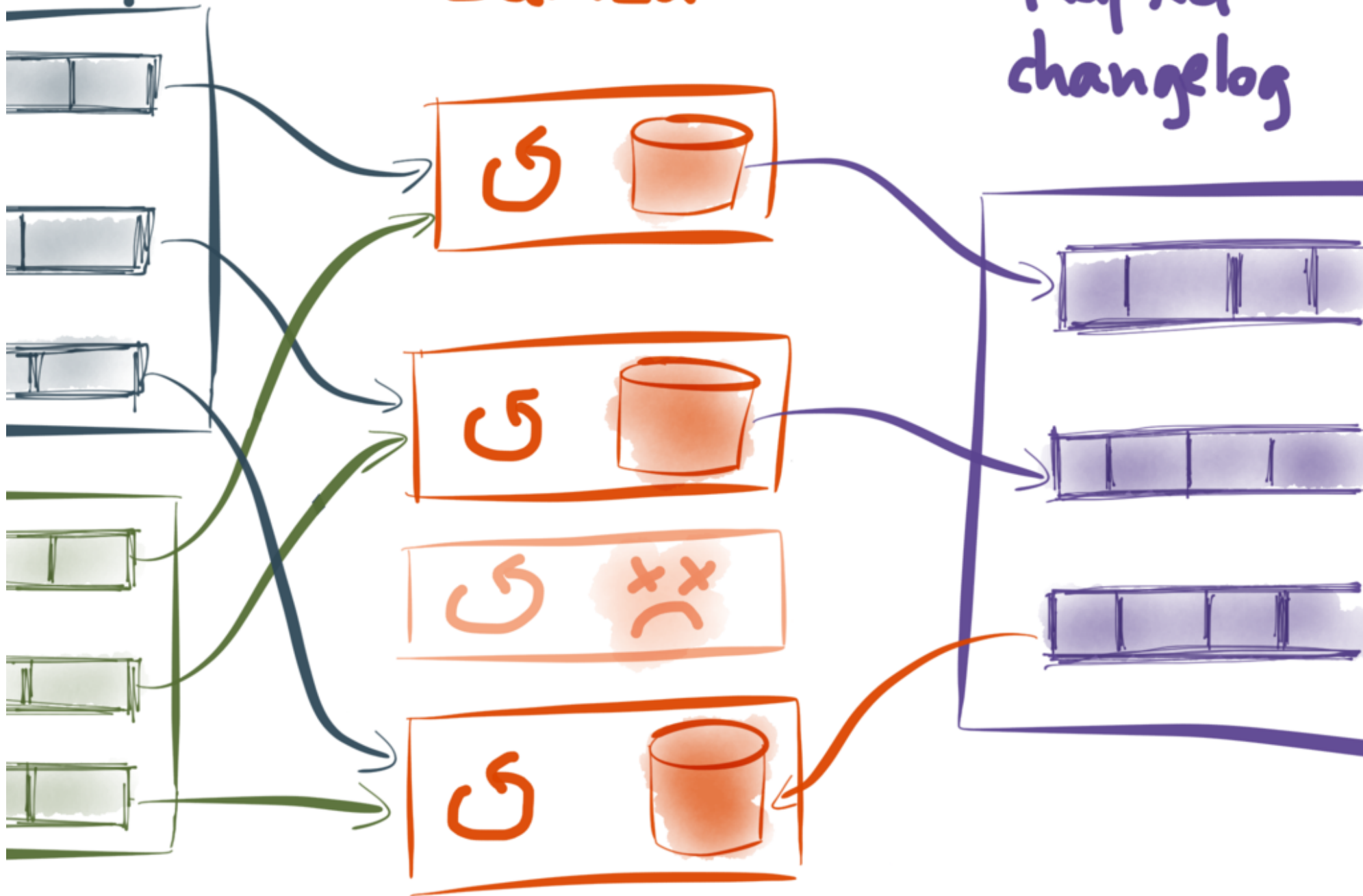
Kafka changelog



Kafka

Samza

Kafka changelog





jfokus



Results for jfokus

[Save](#)[Top](#) / [All](#)

1 new result



Charlotta T @nynelkotten · 4m

T-shirts that actually fit on women might not seem important, but it would be an easy way to make us feel welcome and included. [#Jfokus](#)



Patrick Strang @strangnet · 4m

[@Jfokus](#) true, however since the speed of the queue was so low the serving staff should have been able to notice when plates were emptying

[View conversation](#)

Martin Kleppmann @martinkl · 5m

Hello [#Jfokus](#)! I'll be speaking about scalable real-time data processing with [@samzastream](#), today 14:00 in room A4. martinkl.com/jfokus



Log file

Append



yesterday



last hour



last
minute



jfokus



Results for jfokus

[Save](#)[Top](#) / [All](#)[1 new result](#)

Charlotta T @nynelkotten · 4m

T-shirts that actually fit on women might not seem important, but it would be an easy way to make us feel welcome and included. [#Jfokus](#)



Patrick Strang @strangnet · 4m

[@Jfokus](#) true, however since the speed of the queue was so low the serving staff should have been able to notice when plates were emptying

[View conversation](#)

Martin Kleppmann @martinkl · 5m

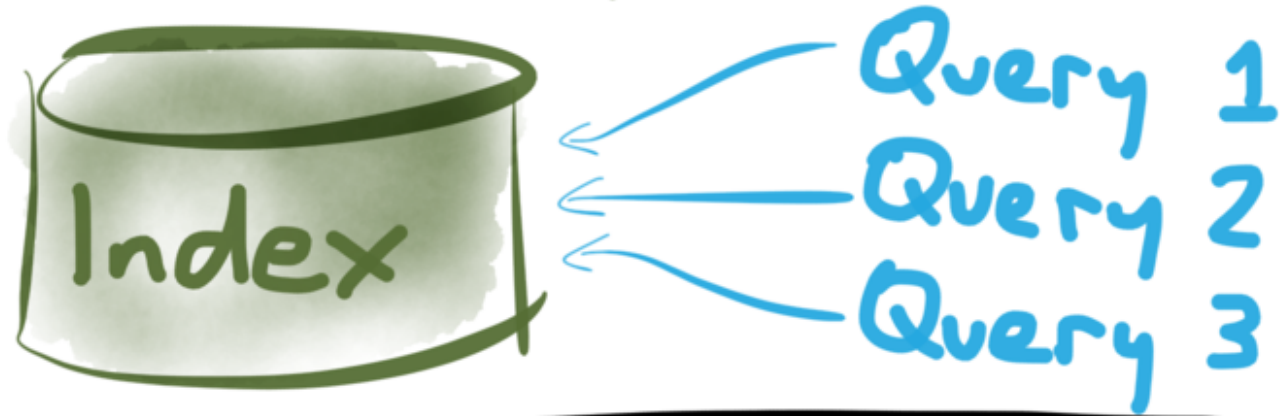
Hello [#Jfokus](#)! I'll be speaking about scalable real-time data processing with [@samzastream](#), today 14:00 in room A4. martinkl.com/jfokus



① After-the-fact search

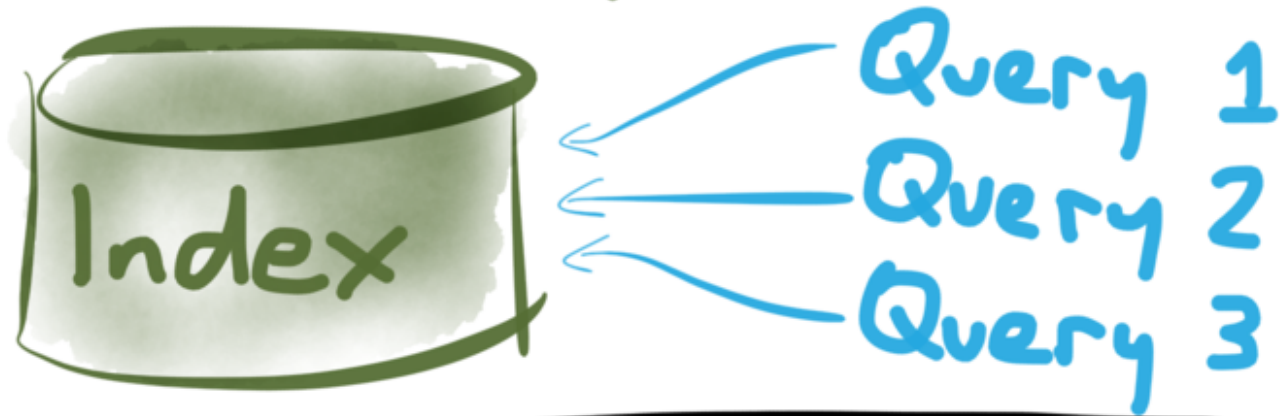
② Streaming search

① After-the-fact search



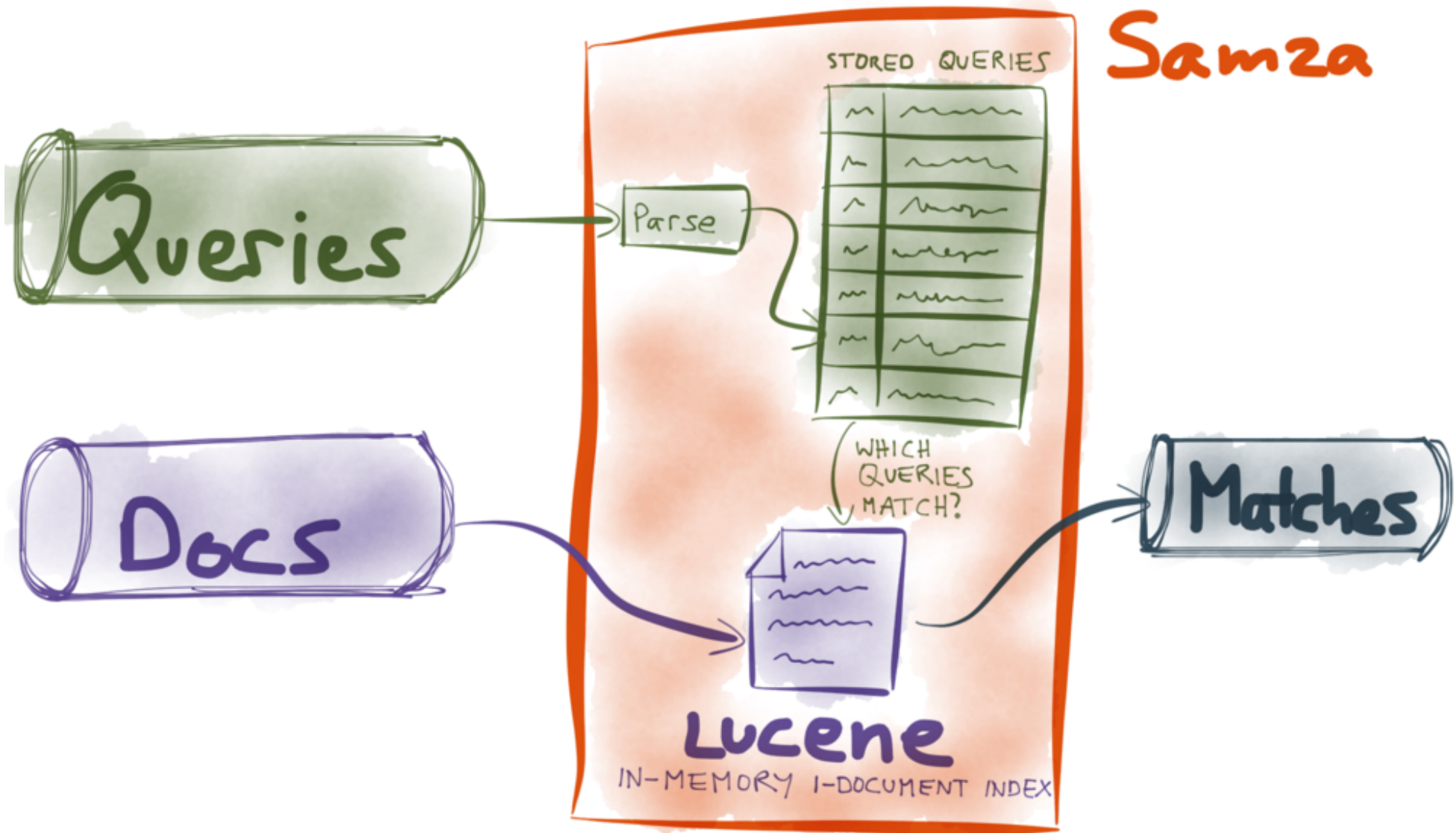
② Streaming search

a) After-the-fact search



b) Streaming search





github.com/romseygeek/samza-luwak

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. 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/slides-das.pdf>
9. Apache Samza documentation. <http://samza.incubator.apache.org>
10. Alan Woodward and Martin Kleppmann: “Samza-Luwak Proof of Concept.” 10 November 2014. <https://github.com/romseygeek/samza-luwak>

dataintensive.net



@martinkl

samza.apache.org

samza