

Apache Kafka Sydney Meetup

That looks weird!

Exploring Mastodon user
behaviour with Kafka & DuckDB

 data-folks.masto.host/@saubury

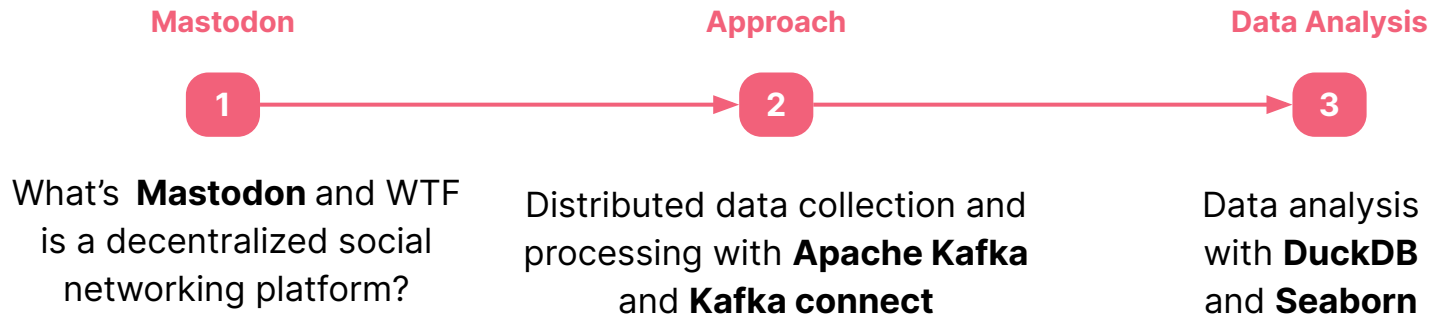
 @SimonAubury

 @saubury



Mastodon user behaviour

What are we talking about today?



Simon Aubury

Principal Data Engineer

/thoughtworks



Kafka enthusiast



Confluent Community Catalyst



Sydney, Australia



Mastodon



Mastodon

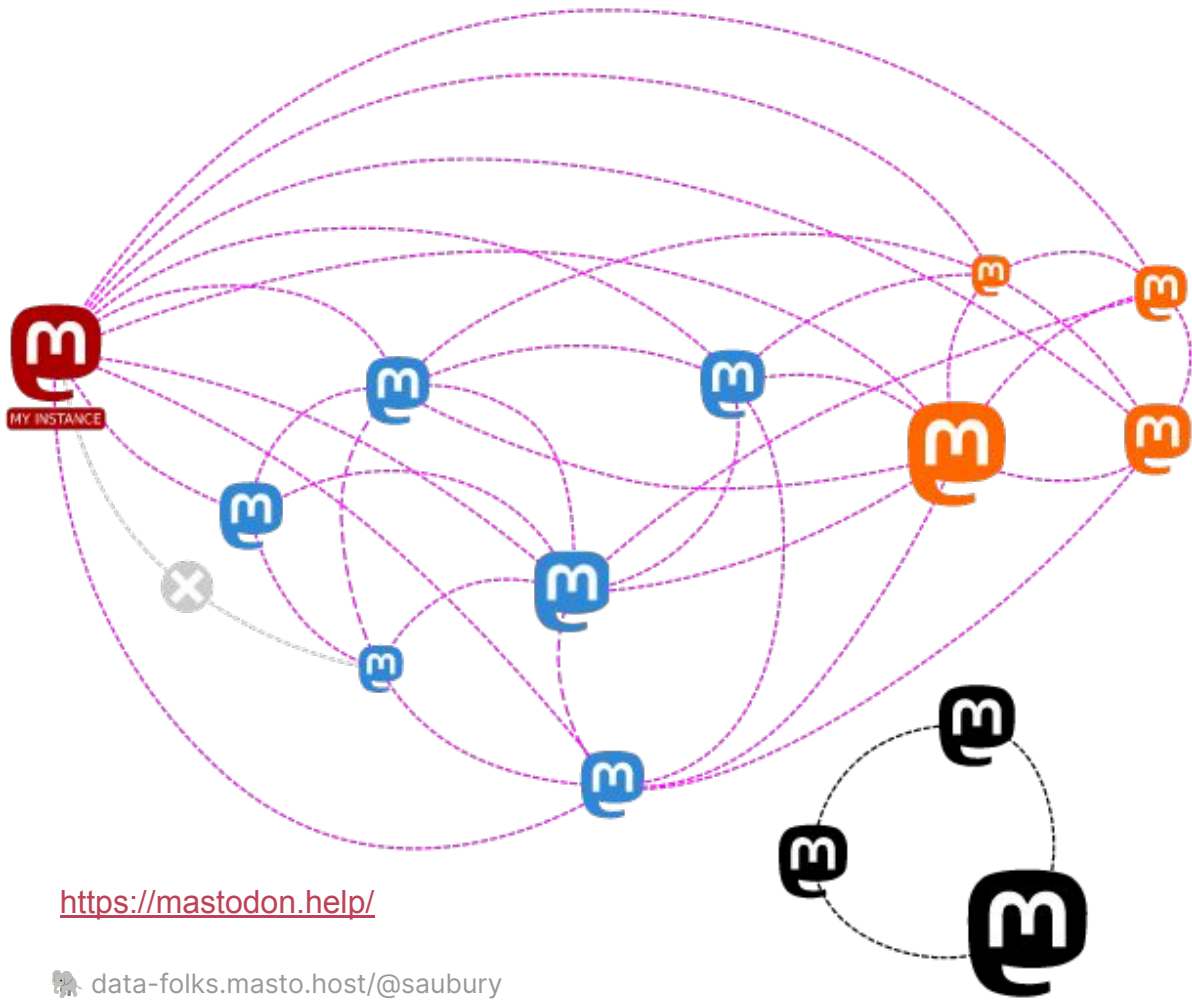
Mastodon is a *decentralized* social networking platform.

- Users are members of a **specific Mastodon instance**
- Servers are capable of joining other servers to **form a federated social network**.

Great for
independent
communities

Difficult to
get global
themes



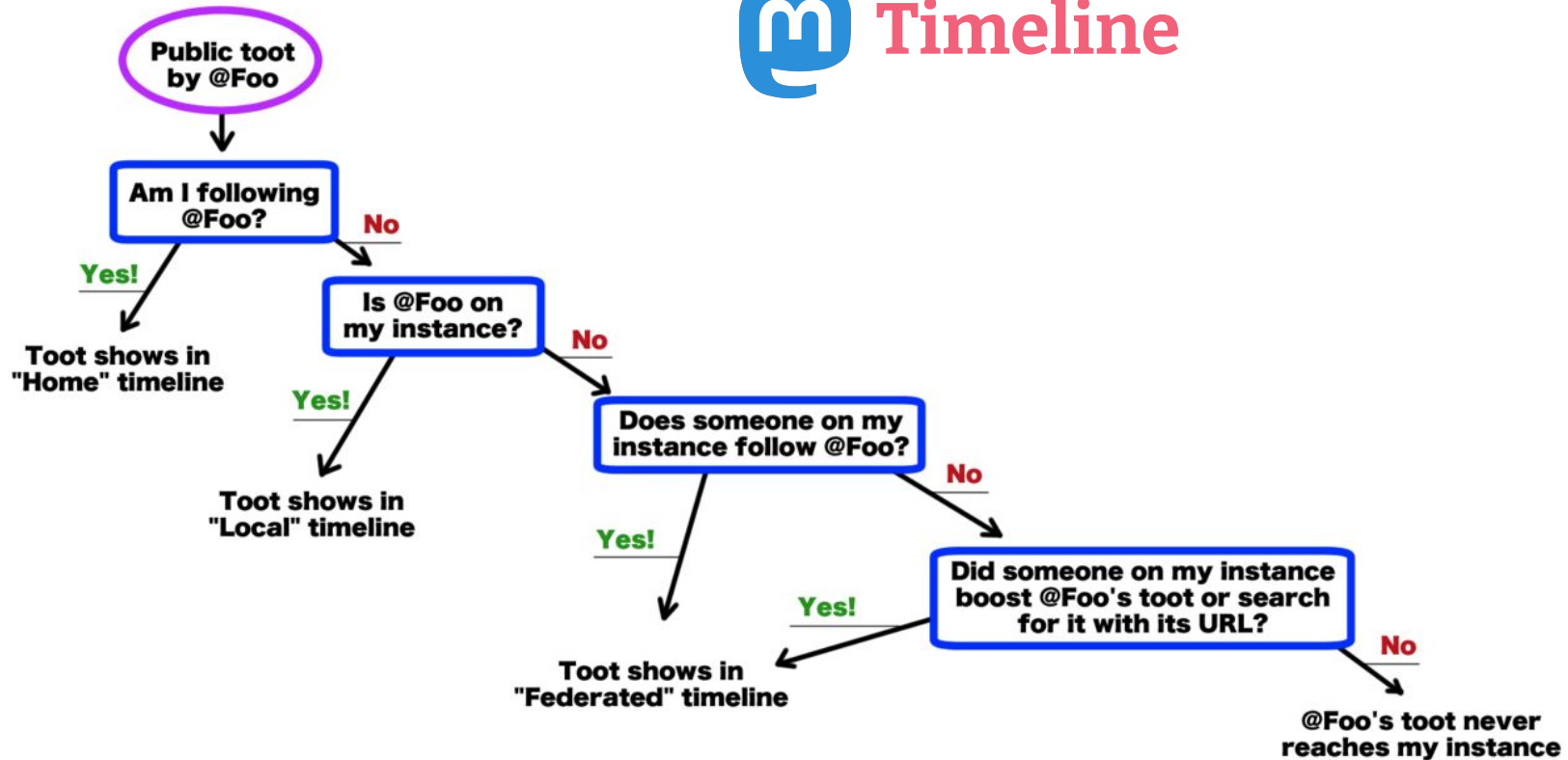


<https://mastodon.help/>

 data-folks.masto.host/@saubury

Mastodon is a galaxy of decentralized and independent networks called instances ...

Each one with its own website, rules and community.



[Wikipedia](https://en.wikipedia.org/wiki/Mastodon)

Data collection



data-folks.masto.host/@saubury

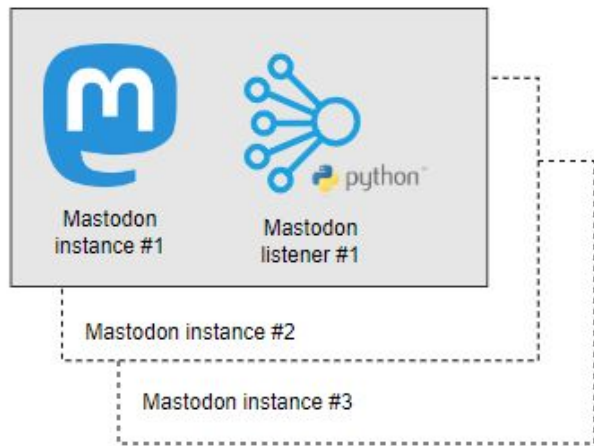


Image by [Pexels](#) from [Pixabay](#)

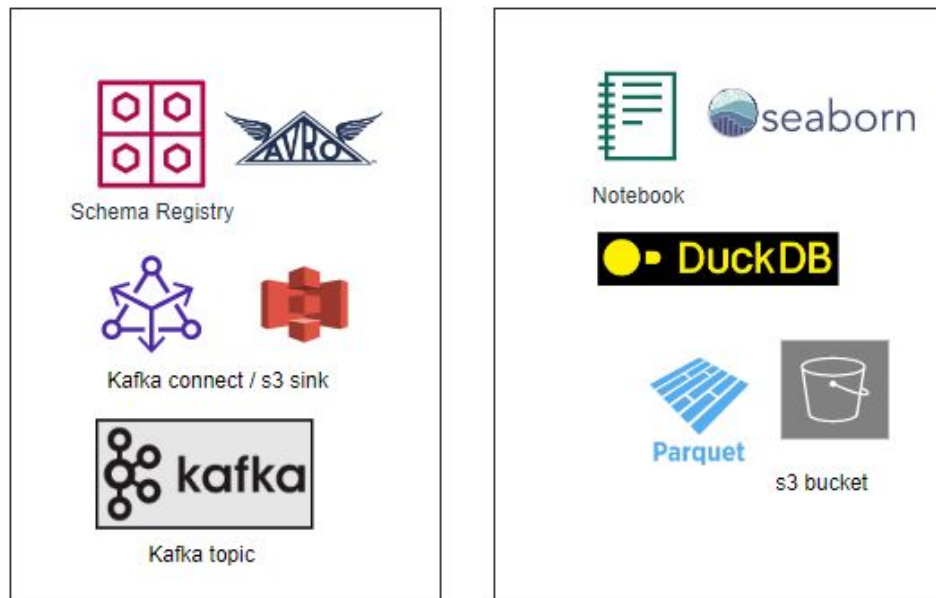
Architecture



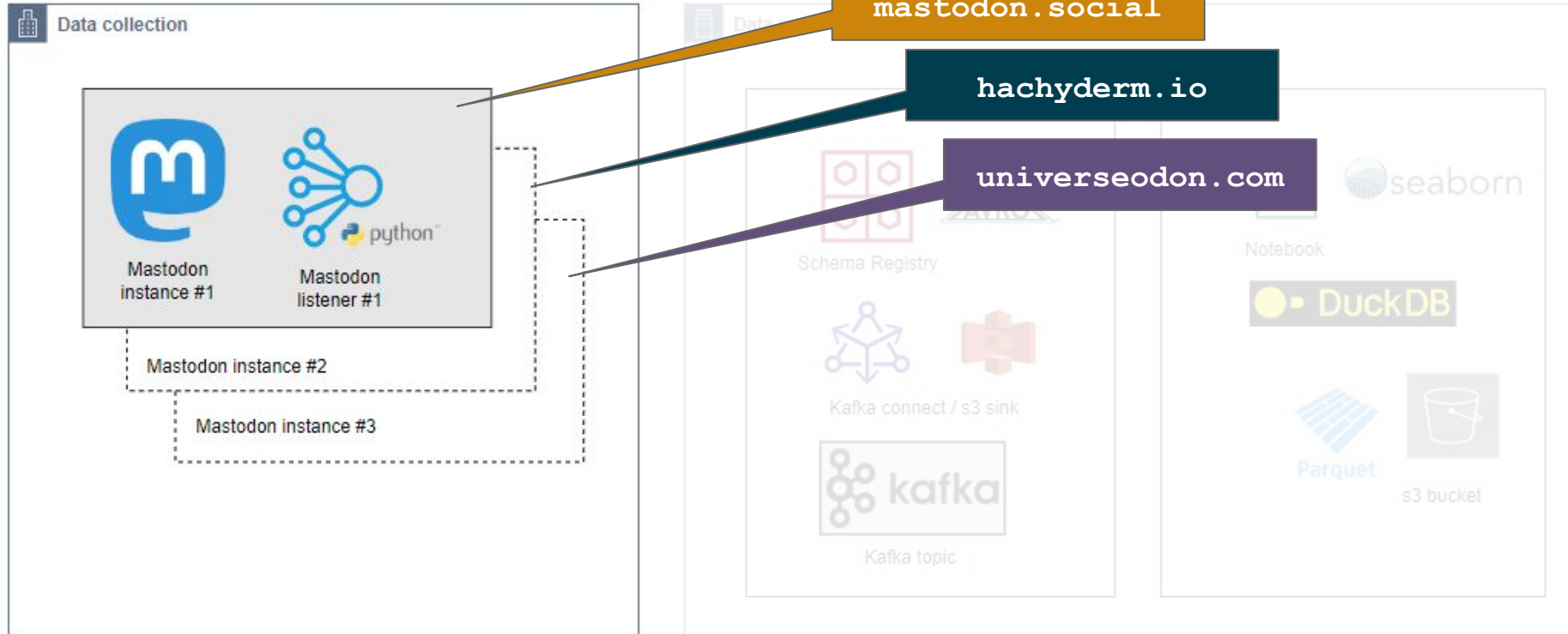
Data collection



Data processing



Data collection





Mastodon Listener

Kafka producer

```
# Kafka AVRO producer
def kafka_producer():
    producer_config = {
        'bootstrap.servers': 'localhost:9092',
        'schema.registry.url': 'http://localhost:8081',
        'broker.address.family': 'v4'
    }
    value_schema = avro.load('avro/mastodon-topic-value.avsc')
    producer = AvroProducer(producer_config, default_value_schema=value_schema)
```

AVRO serializer

Listener

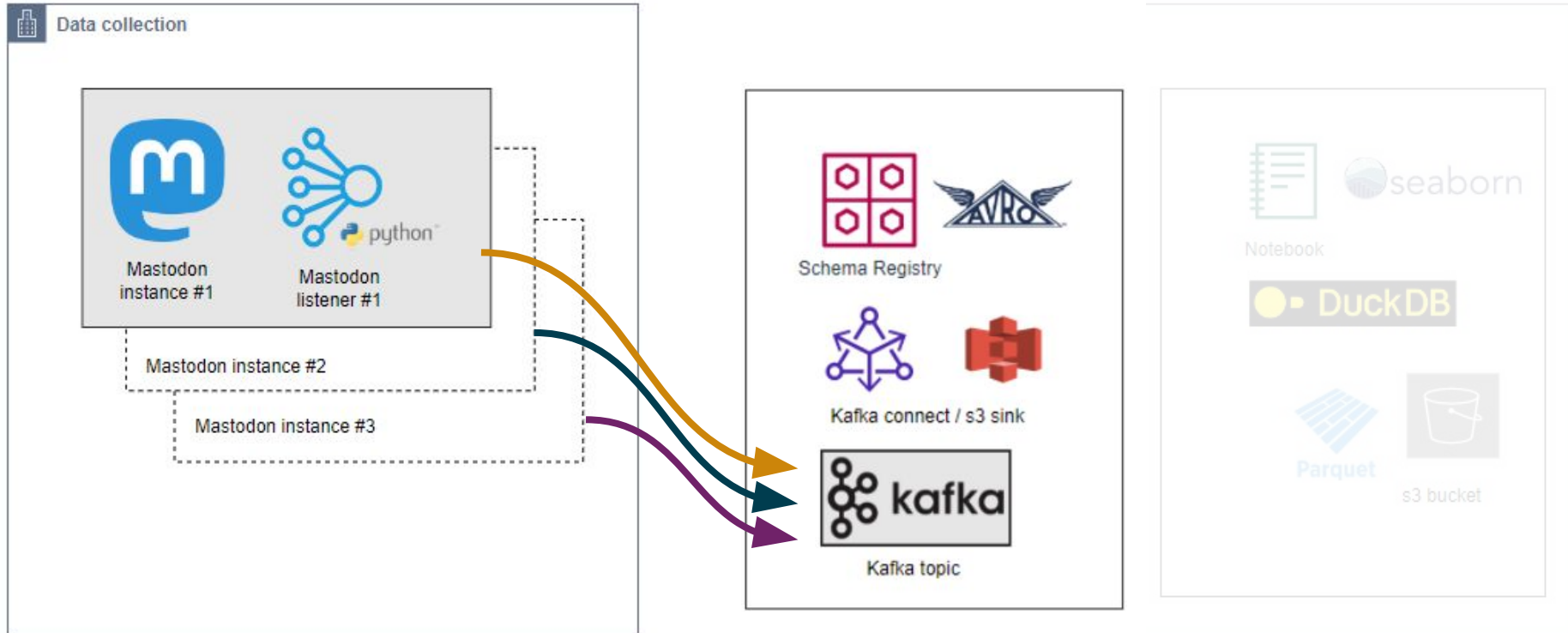
```
# Listener for Mastodon events
class Listener(mastodon.StreamListener):
    def on_update(self, status):
        m_text = BeautifulSoup(status.content, 'html.parser').text
        producer.produce(topic = topic_name, value = value_dict)
```

Servers

```
mastodon = Mastodon('https://mastodon.social')
mastodon.stream_public(Listener())
```



Streaming - Kafka



Kafka - CCC

```
Python
bulwarkonline en "No shortage of Republicans ha
dnddeutsch de #LarianStudios hat den Release
RSMackInnon en b10 EWE: UN biodiversity summi
realTuckRumper en Trump allies demand purge of J
AgentStrangeTV en Why does Miles Morales run war
alliebgamersall en Odious nonsense. Labour cannot
PaulKlee en Magdalena before the conversio
liquid_sunray de Nein, ich gucke nicht die erst
ShawnMcCarthy en @elonmusk Thank you fake Elon!
gmcatch en You know that a government is
WillyECoyote69 en Eigenlijk wil ik knuffelen, ma
cgi unknown Esta noticia me causa una gran
loutre en going to school only 6h or 7h
yantou zh 科普上的评论都好好笑呀
tugu ja 「世の中には性的消費されるために股どい格好してる人もいるし性
Caladin en bro some of y'all's toots need
LucianFreud en The Butcher's Daughter, 2000 #
limbot_gyx en #MEweather #Mw41 ACUS11 KRW
RedditGoneWild en I work with heavy equipment #W
PROSPBerwick en #ASA294 / A4DF56: Squawk 3311,
taniel en Chicago has created new electe
RSMackInnon en b10 EWE: Rook and swift added
toonces en My kid: You know that expressi
crashkrissy en Today's accomplishments: dropp
karasu_sue ja @dax_online Lemmyへの投稿例https://
[]

Python
Shoro fr Vehicle emissions may cause ov
gratefulread en Jail deaths hit 20-year high,
Malik de #mark I am not sure about my S
TsukiokaToshitashi en Tsukioka Yoshitashi, Appearing
nasc en #Earthquake (#aisma) #S.9 stri
SavaFax en New pet, who is this? https://
darkosubotica en So, a while back I've manage t
jikadeau en Sushi restaurants in Japan hav
davebank en @ebutterfly
siguza en (then again, that would be com
Gondathlin es Veré el primer cap de Carnival
dkedrosky en Somewhere over Nevada with @lu
ngorbis en "Situating discretion is the gl
NZPriest en @kairysdal Thank you for the
news en ASX edges up as nery Wall Str
Janantos en The Republic of Estonia celeb
rsfpau en Has anyone asked #ChatGPT to w
artefactCollection en @obekil Tepe is the oldest tem
ling en New local transit survey! A hi
bulwarkonline en In California, the wine and we
bulwarkonline en "No shortage of Republicans ha
PaulKlee en Magdalena before the conversio
feijoa en Doing the #cancer shuffle toda
WillyECoyote69 en Eigenlijk wil ik knuffelen, ma
toonces en My kid: You know that expressi
[]

Python
roya en Signage at the dialysis clinic
siguza en (then again, that would be com
Independent en Trump ordered to give evidence
hoffm en Do we report the height and we
shamui en It's a niche question so yes,
begomit en ARTICLE 1. NAME OF ORGANIZATIO
ngorbis en "Situating discretion is the gl
jeffsheets en I asked ChatGPT if people trus
Sivation en Finally got all of garmentcraf
drclareharris en With more snow on the way... how
uavideos en t.me/ukratna.novosti/50662
NZPriest en @kairysdal Thank you for the
wolfe en i wish someone would love me a
techbolt en #waffle398 5/5
wynch en ALL BLACK People Should Be Awa
bulwarkonline en In California, the wine and we
malcontato pt Visualizei um corpo reclamando
bulwarkonline en "No shortage of Republicans ha
dnddeutsch de #LarianStudios hat den Release
alliebgamersall en Odious nonsense. Labour cannot
PaulKlee en Magdalena before the conversio
loutre en going to school only 6h or 7h
Caladin en bro some of y'all's toots need
NormanRockwell en Babysitter, 1927 #regionalism
toonces en My kid: You know that expressi
[]

Python
bulwarkonline en "No shortage of Republicans ha
dnddeutsch de #LarianStudios hat den Release
RSMackInnon en b10 EWE: UN biodiversity summi
realTuckRumper en Trump allies demand purge of J
AgentStrangeTV en Why does Miles Morales run war
alliebgamersall en Odious nonsense. Labour cannot
PaulKlee en Magdalena before the conversio
liquid_sunray de Nein, ich gucke nicht die erst
ShawnMcCarthy en @elonmusk Thank you fake Elon!
gmcatch en You know that a government is
WillyECoyote69 en Eigenlijk wil ik knuffelen, ma
cgi unknown Esta noticia me causa una gran
loutre en going to school only 6h or 7h
yantou zh 科普上的评论都好好笑呀
tugu ja 「世の中には性的消費されるために股どい格好してる人もいるし性
Caladin en bro some of y'all's toots need
LucianFreud en The Butcher's Daughter, 2000 #
limbot_gyx en #MEweather #Mw41 ACUS11 KRW
RedditGoneWild en I work with heavy equipment #W
PROSPBerwick en #ASA294 / A4DF56: Squawk 3311,
taniel en Chicago has created new electe
RSMackInnon en b10 EWE: Rook and swift added
toonces en My kid: You know that expressi
crashkrissy en Today's accomplishments: dropp
karasu_sue ja @dax_online Lemmyへの投稿例https://
[]
```

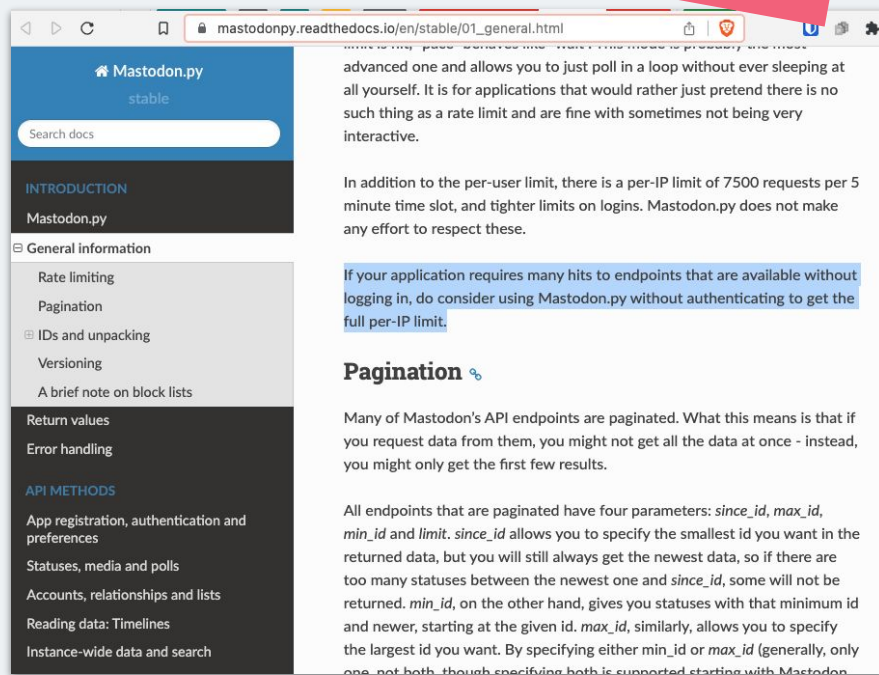
Authenticating

Mastodon.py

You generally *don't* need your application to be authorized to listen to a mastodon feed.



The docs recommend to ***not*** authenticate



The screenshot shows the Mastodon.py documentation website. The browser address bar displays `mastodonpy.readthedocs.io/en/stable/01_general.html`. The page header includes the Mastodon.py logo and the word 'stable'. A search bar is present. The sidebar on the left contains a list of navigation links: 'INTRODUCTION', 'Mastodon.py', 'General information' (expanded), 'Rate limiting', 'Pagination', 'IDs and unpacking', 'Versioning', 'A brief note on block lists', 'Return values', 'Error handling', 'API METHODS', 'App registration, authentication and preferences', 'Statuses, media and polls', 'Accounts, relationships and lists', 'Reading data: Timelines', and 'Instance-wide data and search'. The main content area on the right discusses rate limiting and pagination. A pink callout box at the top right states: 'The docs recommend to ***not*** authenticate'. The text in the main content area includes: 'advanced one and allows you to just poll in a loop without ever sleeping at all yourself. It is for applications that would rather just pretend there is no such thing as a rate limit and are fine with sometimes not being very interactive.' and 'In addition to the per-user limit, there is a per-IP limit of 7500 requests per 5 minute time slot, and tighter limits on logins. Mastodon.py does not make any effort to respect these.' A blue highlighted box contains the text: 'If your application requires many hits to endpoints that are available without logging in, do consider using Mastodon.py without authenticating to get the full per-IP limit.'

Pagination

Many of Mastodon's API endpoints are paginated. What this means is that if you request data from them, you might not get all the data at once - instead, you might only get the first few results.

All endpoints that are paginated have four parameters: `since_id`, `max_id`, `min_id` and `limit`. `since_id` allows you to specify the smallest id you want in the returned data, but you will still always get the newest data, so if there are too many statuses between the newest one and `since_id`, some will not be returned. `min_id`, on the other hand, gives you statuses with that minimum id and newer, starting at the given id. `max_id`, similarly, allows you to specify the largest id you want. By specifying either `min_id` or `max_id` (generally, only one, not both, though specifying both is supported starting with Mastodon

Data storage



data-folks.masto.host/@saubury

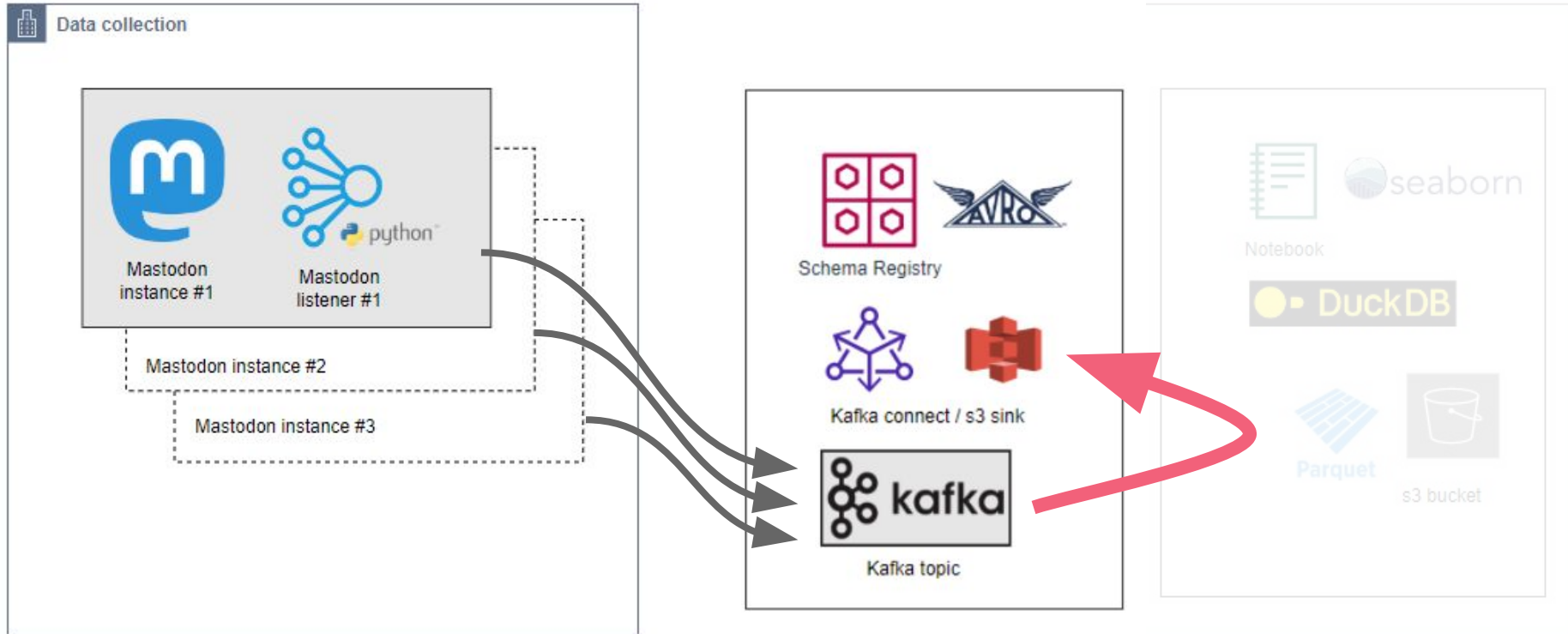


Image by [Pexels](#) from [Pixabay](#)

Data storage



Store to s3



Kafka to s3

Kafka connect / s3 sink

Connector name

S3 sink class

Kafka topic

Write as parquet

S3 bucket details

```
{  
  "name": "mastodon-sink-s3-aws",  
  "connector.class": "io.confluent.connect.s3.S3SinkConnector",  
  "topics": "mastodon-topic",  
  "format.class": "io.confluent.connect.s3.format.parquet.ParquetFormat",  
  "flush.size": "1000",  
  "s3.bucket.name": "2023mastodon",  
  "aws.access.key.id": "XXxxXXxxXX",  
  "aws.secret.access.key": "XXxxXXxxXXXXxxXXxxXXXXxxXXxxXXXXxxXXxxXX",  
  "s3.region": "us-east-1",  
  "storage.class": "io.confluent.connect.s3.storage.S3Storage"  
}
```



DuckDB



data-folks.masto.host/@saubury

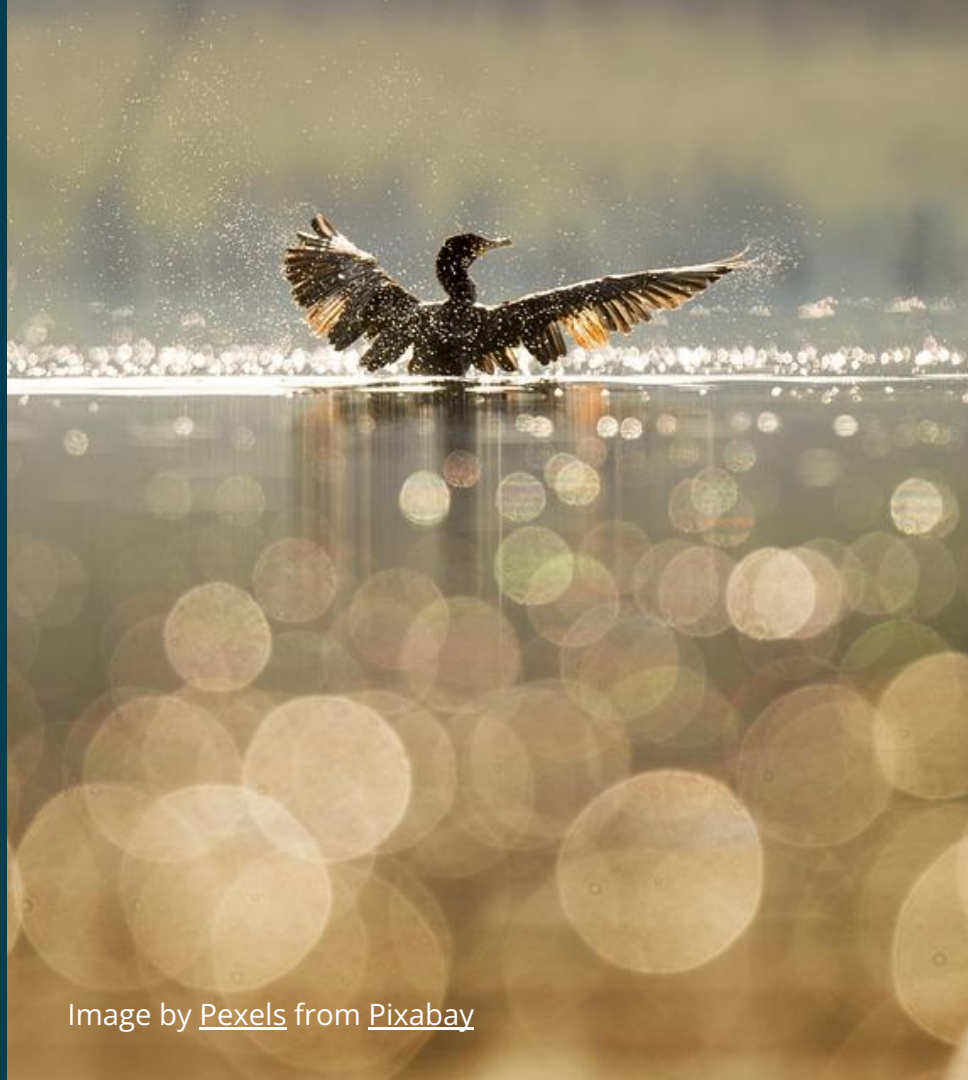
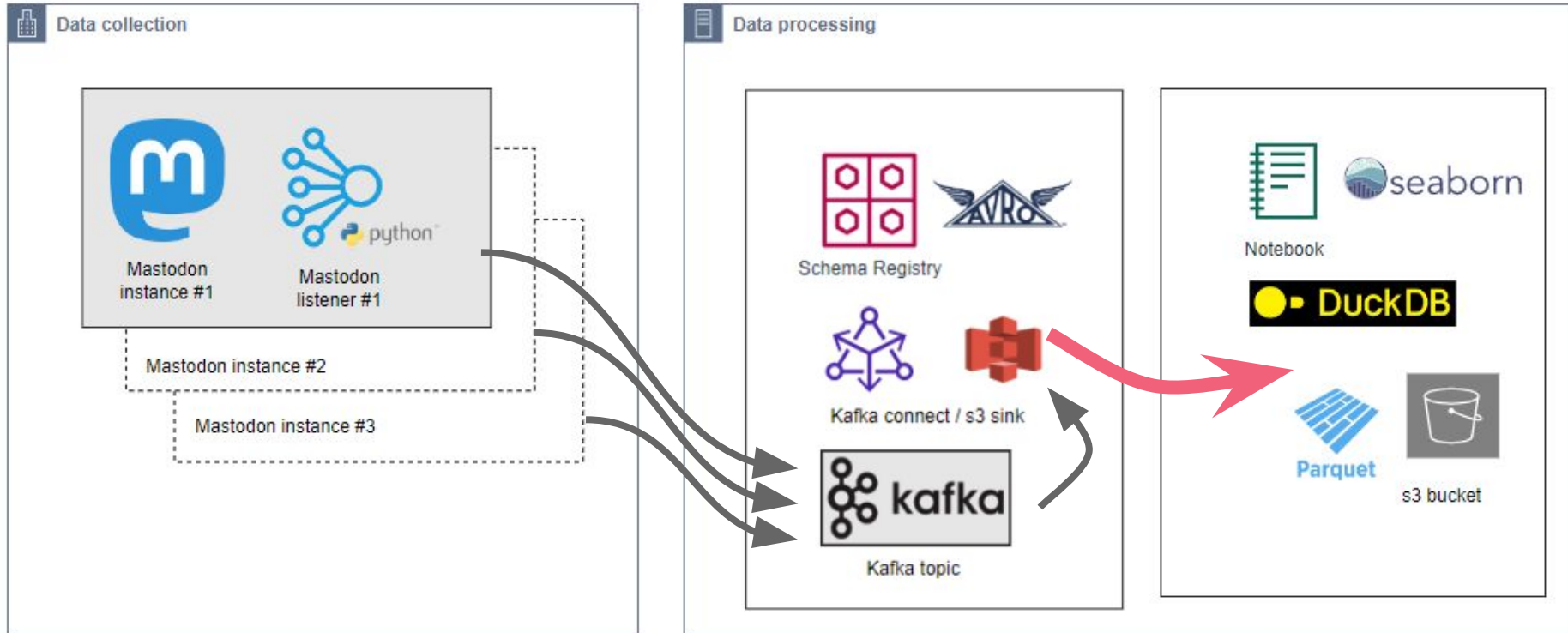


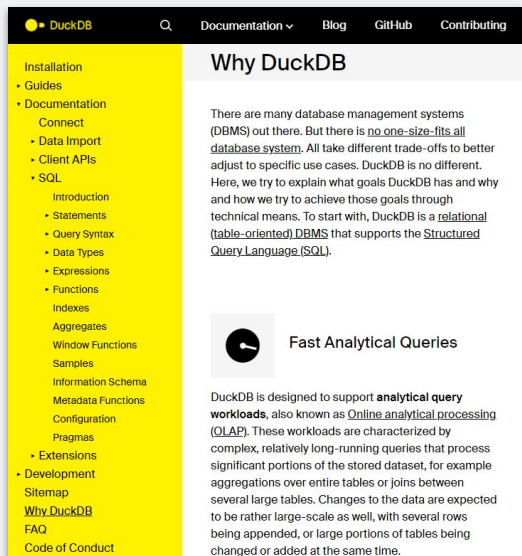
Image by [Pexels](#) from [Pixabay](#)

Parquet in DuckDB



Why DuckDB

DuckDB is free / open-source



https://duckdb.org/why_duckdb.html



RDBMS / SQL / OLAP - Extensive SQL with a large function library, window functions etc. ACID transactional



No external dependencies - compiled into two files, embedded within a host process. DuckDB uses **PostgreSQL's SQL parser**, Google's RE2 regular expression engine and SQLite's shell

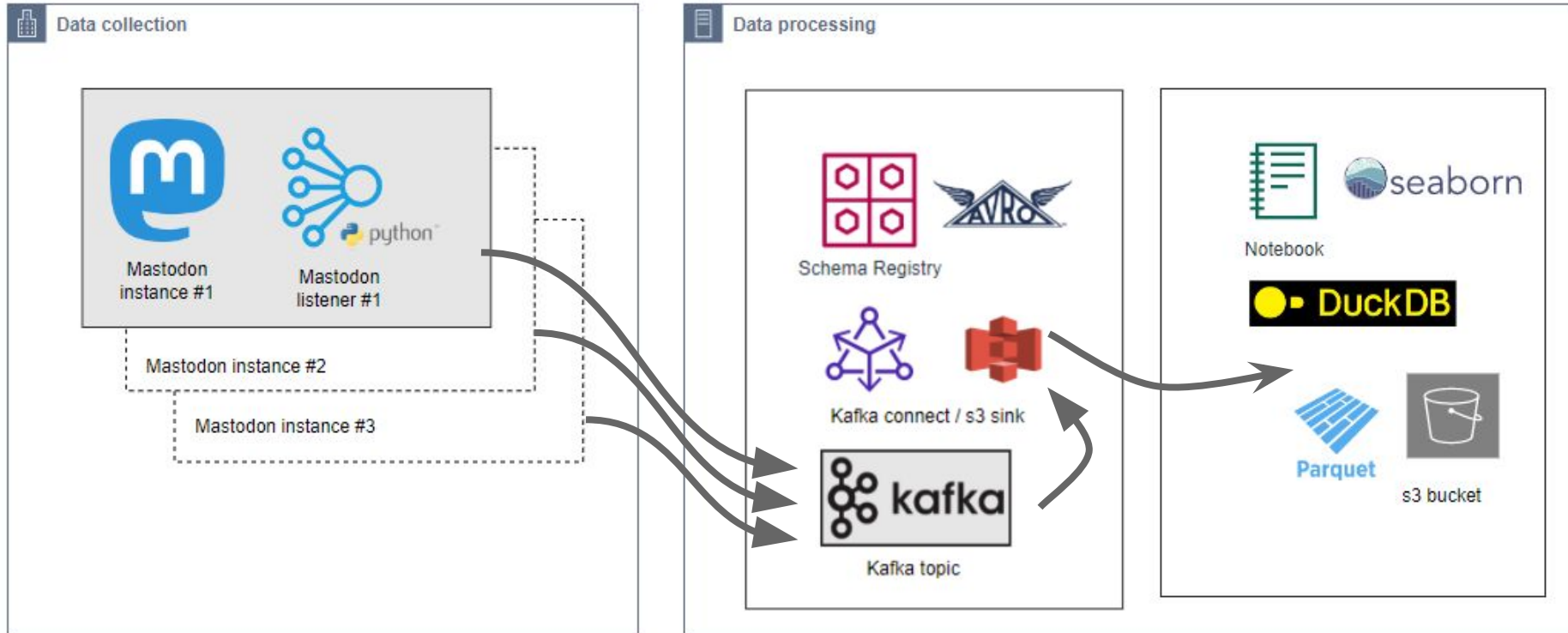


Process foreign data without copying, run queries directly on Pandas data, Python and R, APIs for Java, C, C++



Extensions - Httpfs, s3, parquet, json, FTS, geospatial

Parquet in DuckDB



DuckDB SQL

Start DuckDB

```
./duckdb
```

CLI / macOS

Install extensions

```
INSTALL httpfs;  
LOAD httpfs;
```

Set s3 endpoint

```
set s3_access_key_id='XXxxXXxx';  
set s3_secret_access_key='XXxxXXxxXXxxXXxxXXxxXXxx';  
set s3_region='us-east-1';
```

Select parquet

```
select *  
from  
read_parquet('s3://mastodon/mastodon-topic/partition=0/*');
```



DuckDB SQL

1.6 million

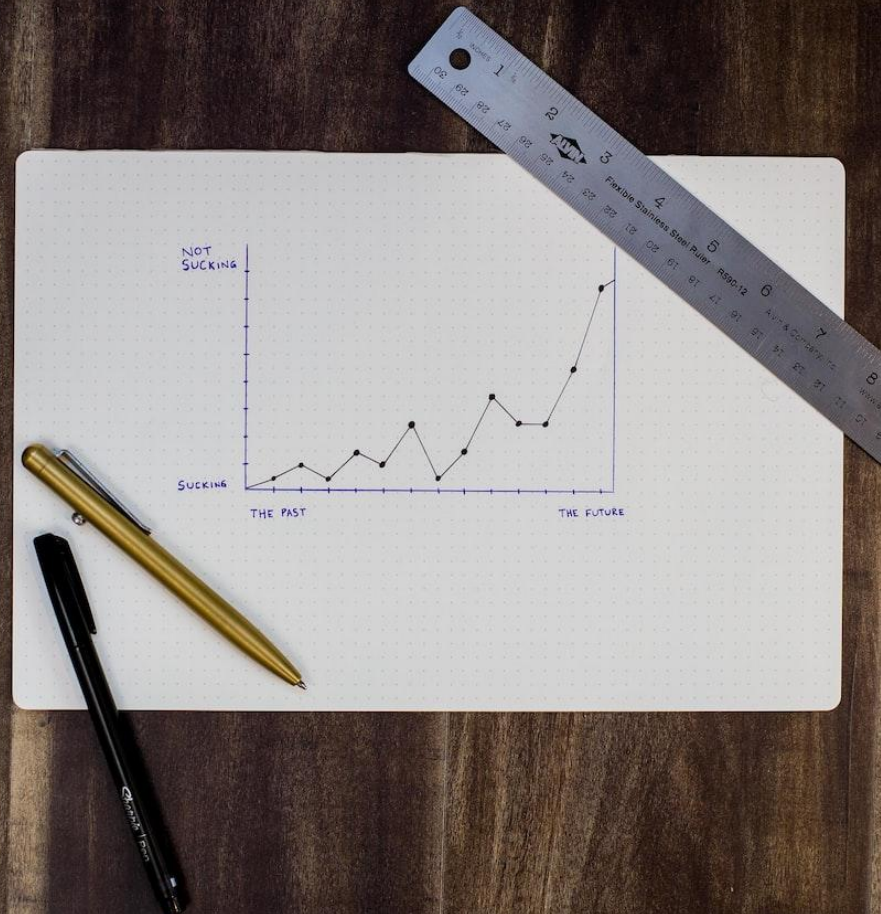
```
create table mastodon_toot
as
select m_id, created_at, app, url, username, mastodon_text
from read_parquet('s3://mastodon-topic/partition=0/*');
```

SQL



2]	✓	0.8s
	Count	
	0	1622149

Data Analysis



Daily Mastodon Usage

```
select strftime(created_tz, '%Y/%m/%d %a')
, count(*) as "Num toots"
, count(distinct(username)) as "Num users"
, count(distinct(from_instance)) as "Num urls"
, mode(case when bot='False' then username end)
, mode(case when bot='True' then username end)
, mode(base_url) as "Most freq host"
from mastodon_toot
group by 1 order by 1;
```

SQL / Notebook

Created day	Num toots	Num users	Num urls	Most freq non-bot	Most freq bot	Most freq host
2023/02/03 Fri	17880	8537	1524	gnutiez	nieuws	https://mastodon.social
2023/02/04 Sat	210646	54006	4562	gnutiez	cnexnews	https://mastodon.social
2023/02/05 Sun	191391	49241	4310	IzumiHal	ua	https://mastodon.social
2023/02/06 Mon	41632	17846	2255	gnutiez	nieuws	https://mastodon.social
2023/02/07 Tue	99097	30701	3350	gnutiez	cnexnews	https://mastodon.social
2023/02/08 Wed	188503	49649	4372	gnutiez	cnexnews	https://mastodon.social
2023/02/09 Thu	166096	48532	4227	worldeconomicfella	cnexnews	https://mastodon.social
2023/02/10 Fri	207877	54230	4608	gnutiez	cnexnews	https://mastodon.social

200,000 toots a day
from 50,000 users,
4,000 servers

mastodon.social was the
most popular host

News organisations are
the biggest generator of
content

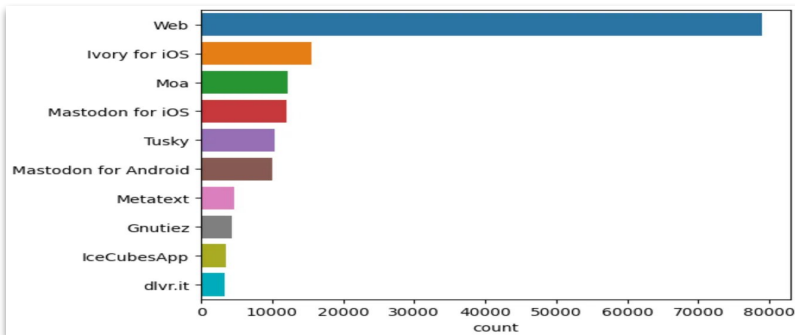


Mastodon App Landscape

```
%%sql
mastodon_app_df <<
  select *
  from mastodon_toot
  where app is not null
  and app <> ''
  and bot='False';

sns.countplot(data=mastodon_app_df, y="app")
```

SQL / Notebook



Mastodon application landscape is rapidly changing

Web usage is the preferred client, with mobile apps like Ivory, Moa, Tusky, and the Mastodon app

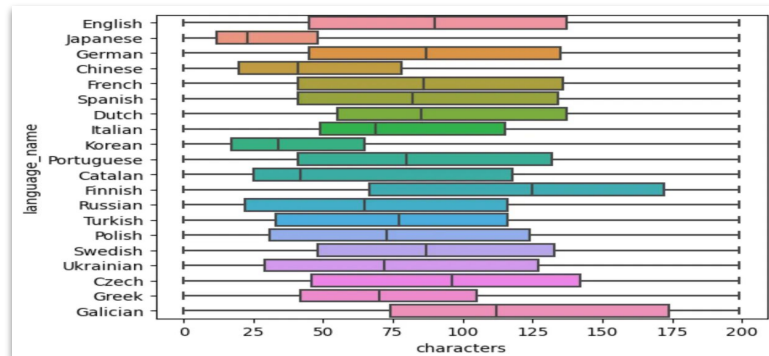
Generally, the app attribute does not federate

Toot Length by Language Usage

```
%%sql
mastodon_lang_df <<
  select *
  from mastodon_toot
  where language not in ('unknown');

sns.boxplot(data=mastodon_lang_df,
x="characters", y="language_name", orient="h")
```

SQL / Notebook



Chinese, Japanese, and Korean toots are shorter than English

Galicia and Finnish messages are longer

Perhaps logographic languages (like Mandarin) convey more with fewer characters?

Longest toot ...

```
%%sql
with cte as
(
    select username
    , mastodon_text
    , len(mastodon_text) as len_toot
    , max(len(mastodon_text)) over () as max_len_toot
    from mastodon_toot
)
select username, mastodon_text
from cte
where len_toot = max_len_toot;
```

SQL / Notebook

68,991 characters



Marsey the Cat

Surprising statistics

Quick statistics from the data collected over **ten days** (February 3 to February 12)



1,622,149 Mastodon toots seen



142,877 unique Mastodon users



8,309 unique Mastodon instances, **131 languages** seen



Shortest toot is 0 characters, **average toot length is 151** characters, and **longest toot is 68,991** characters



All toots 245,245,677 characters (over 1.6 million toots) in **DuckDB's memory only 745.5MB**



Time it takes to calculate the above statistics in a single SQL query is **0.7 seconds**

NOT SO - Surprising discoveries

Kafka - great for scale & resilience

DuckDB - super flexible tool for EDA

Kafka ecosystem - flexible data distribution

The internet is
[fun | weird]




Thanks / questions?

 [Medium Blog](#)

 [Code](#)

 data-folks.masto.host/@saubury

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

 data-folks.masto.host/@saubury



DATAENGBYTES KEY DATES

MARCH 20TH | *CFP Open*

MAY 31ST | *CFP Closed*

AUGUST 21ST | *Perth
Conference*

AUGUST 25TH | *Sydney
Conference*

AUGUST 29TH | *Brisbane
Conference*

AUGUST 31ST | *Melbourne
Conference*

 [DataEngBytes 2023](#)