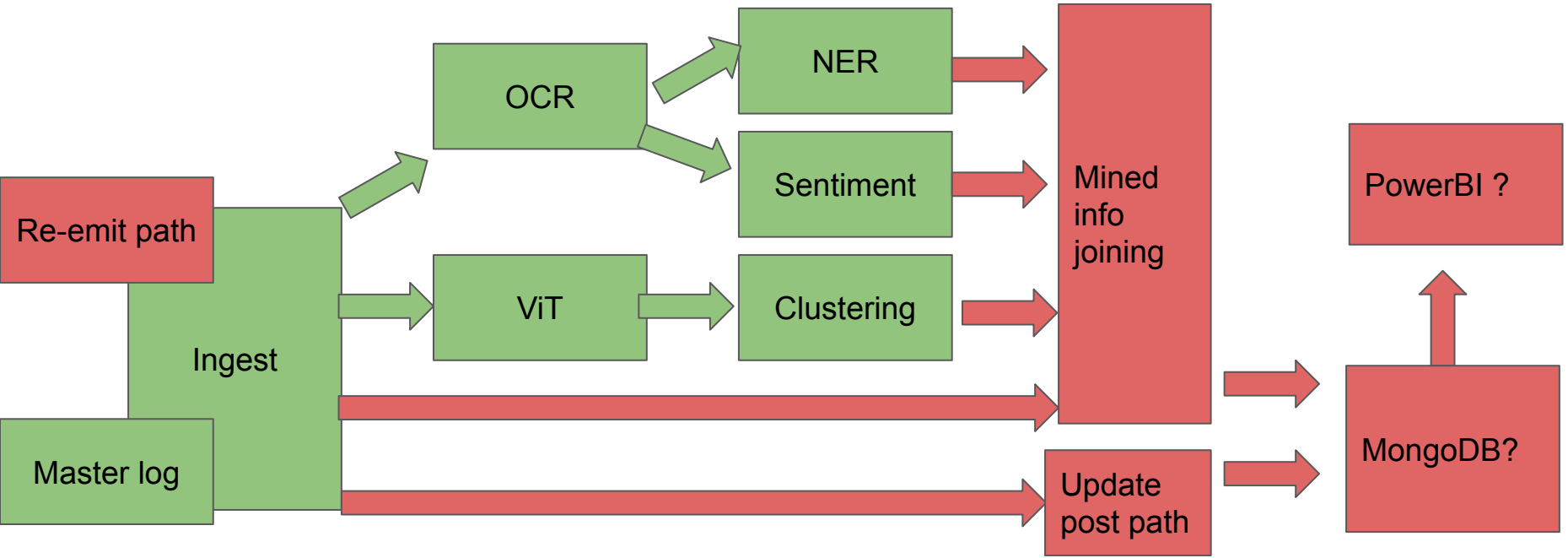
Beeg meme progress

Beeg Data team


Re-emit path
Ingest
Master log
OCR
ViT
NER
Sentiment
Clustering
Mined info joining
Update post path
MongoDB?
PowerBI ?

Kafka queues

- Created a kafka topic for each connection
  - Probably should’ve used consumer groups and so on
  - But this is much simpler and works for our use case
- Using a script
- Might be a better way
- Again, this is simpler

```
h --bootstrap-server localhost:9092 --create --if-not-exists --topic "nifi2vit"
h --bootstrap-server localhost:9092 --create --if-not-exists --topic "nifi2ocr"
h --bootstrap-server localhost:9092 --create --if-not-exists --topic "nifi2analytics"
h --bootstrap-server localhost:9092 --create --if-not-exists --topic "vit2cluster"
h --bootstrap-server localhost:9092 --create --if-not-exists --topic "vit2analytics"
h --bootstrap-server localhost:9092 --create --if-not-exists --topic "cluster2analytics"
h --bootstrap-server localhost:9092 --create --if-not-exists --topic "ocr2ner"
h --bootstrap-server localhost:9092 --create --if-not-exists --topic "ocr2sentiment"
h --bootstrap-server localhost:9092 --create --if-not-exists --topic "ocr2analytics"
h --bootstrap-server localhost:9092 --create --if-not-exists --topic "ner2analytics"
h --bootstrap-server localhost:9092 --create --if-not-exists --topic "sentiment2analytics"
```

Hacking Structured Streaming

- Spark Structured Streaming not really designed for this
- But has UDF (user-defined function)
  - We'll manage
- No batching unfortunately
  - There might be a better way
  - We wanted to get most of the system *somewhat working* already

```
from sklearn.cluster import Birch

m = Birch()

@udf(IntegerType())
def mult(x, y):
    m.partial_fit([[x, y]])
    return int(m.predict([[x, y]])[0])
```

ViT - image embeddings

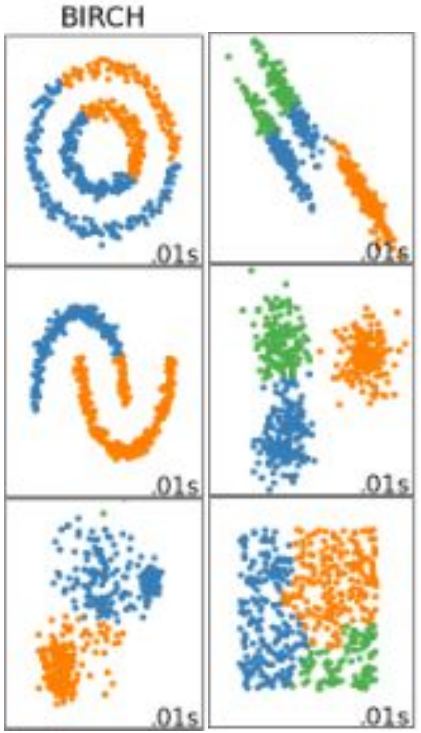
- Used a version of ViT designed for mobile phones
  - Works pretty well, benchmark on right
  - But tested on a local PC, not cloud
  - On cloud it's quite slow (~1 s per image)
    - Probably downloading
    - How do you even profile spark structured streaming?
- Sometimes has issues with inferring image sizes

```
%%timeit

inputs = image_processor(image, return_tensors="pt")

with torch.no_grad():
    outputs = model(**inputs)
```

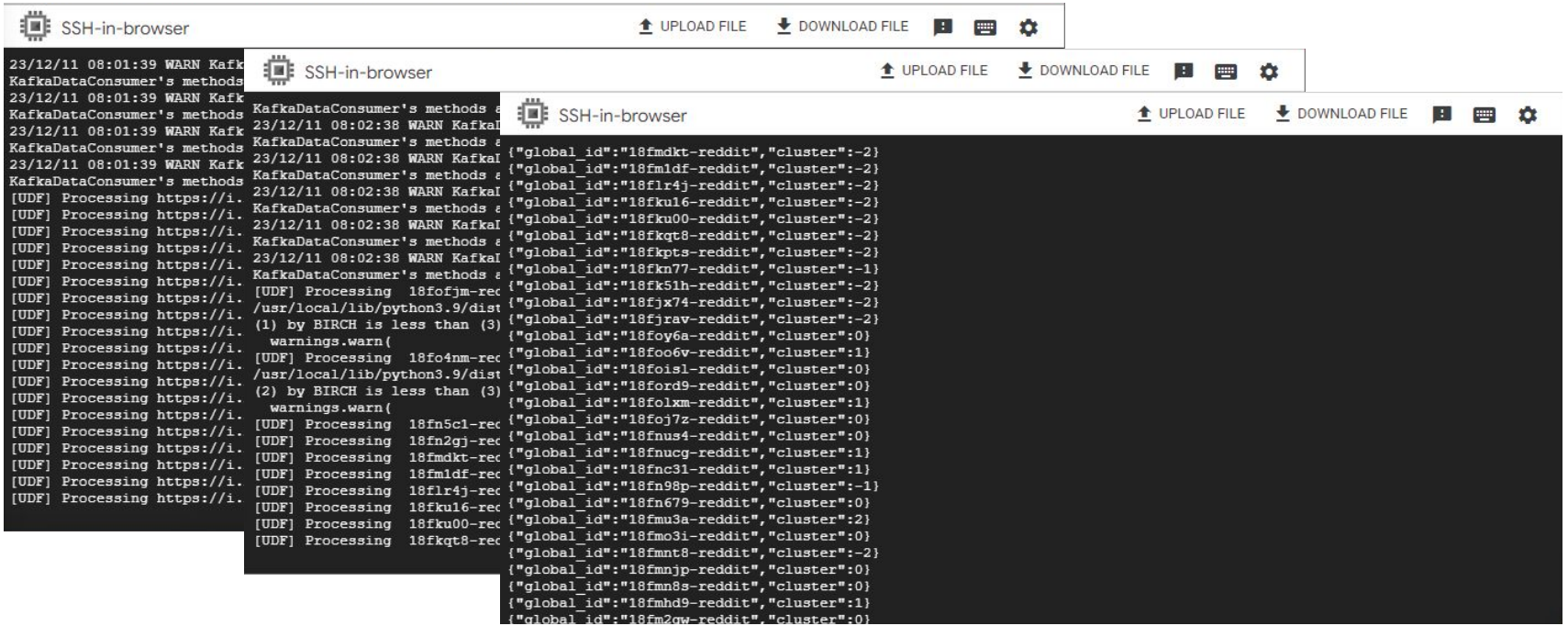
```
63 ms ± 1.32 ms per loop (mean ± std. dev. of 7 runs,
```

Clustering - own model

- Online learning
- Using BIRCH from scikit-learn
- Bare minimum working
  - Some non-error numbers coming out the other end
  - No clue how well it’s working
- Saves model to checkpoints
- Doesn’t load the model from the checkpoints (yet)



Clustering-path testing



Optical Character Recognition

- Used EasyOCR python module
  - Slow on a local PC (CPU)
  - Even slower on the cloud
- Sometimes has issues with a couple of characters in a word

```
model = easyocr.Reader(['en'])

start = time.time()
image = PIL.Image.open(requests.get(url2, stream=True).raw)
open_cv_image = np.array(image.convert('RGB'))
results = model.readtext(open_cv_image)
end = time.time()
print(f"{end - start} seconds")
```

✓ 0.0s

```
11.350 seconds
```


```
['mobile games 20108', 'mobie games 20208']
```

Optical Character Recognition Testing

```
tesliukillia@nifi:~$ sudo su
root@nifi:/home/tesliukillia# cd ../nifi
root@nifi:/home/nifi# kafka_2.13-3.6.0/bin/kafka-console-consumer.sh --bootstrap-server nifi:9092 --topic ocr2analytics
{"global_id":"18frl4v-reddit","text":"\"intel. downloadmoreram free; fast and instant !. amd. \""}
{"global_id":"18fregc-reddit","text":"\"the guy who just left a 12 hour shift that he had to do on the 24.12 to save enough for the proposal ring; but adam sandler was teaching his gf the meaning of christmas the entire time and now she left him. uaudle witl ielaitic. \""}
{"global_id":"18frabr-reddit","text":"\"[ threwup. \""}
{"global_id":"18fr865-reddit","text":"\"water. h2o. hydrogen oxide. made with mematic. \""}
{"global_id":"UfDit3n-imgur","text":""}
{"global_id":"9GdVtUD-imgur","text":""}
{"global_id":"G6Q2ARh-imgur","text":""}
{"global_id":"18frv5d-reddit","text":"\[redacted] \""}
{"global_id":"18fwv3c-reddit","text":""}
{"global_id":"18fr7jk-reddit","text":"\"minecraft fans complaining that updates take too long. gta5 fans waiting for the next game . look what they need to mimic a fraction of our power. \""}
{"global_id":"18fwzfb-reddit","text":"\"\\\"call of duty modern warfare please\\\" befch] chhich] one call um call\\\" duta| call duty modern wapeare miwern warfare modern warfare modern warfare. callof duty. callduty na. modern warfare. callduty min niw a a fa e. imgilip_com. \""}
{"global_id":"hFmL7x7-imgur","text":""}
{"global_id":"ciYRinH-imgur","text":""}
{"global_id":"18fx3z6-reddit","text":""}
{"global_id":"736424265473277952-tumblr","text":"\"its 110? in the shade. tmiswhntther is gren cocofun. \""}
{"global_id":"18fx0bs-reddit","text":"\"how sleep knowing my ass is very on the internet and get to post more of it tomorrow. \""}
{"global_id":"wESSEBO-imgur","text":""}
{"global_id":"jSt4g5W-imgur","text":""}
{"global_id":"18fwsln-reddit","text":"\"me telling my parents that | want to be a gamer. my parents. powered dh intel. esl one a lumpur:. stcplay. 1 |. prtdator. \""}
{"global_id":"736423372160909312-tumblr","text":"\"nairobi 50 direct af j04 be. tena netwoll m-pe kilgoris klassic   sacco saco nairobi direct nai (kisumukat/to narok nairobi. lnc :. safaricom `. \""}
```