Beeg Data - Beeg Meme

Kacper Grzymkowski, Jakub Fołtyn,
Illia Tesliuk, Mikołaj Malec

Goal

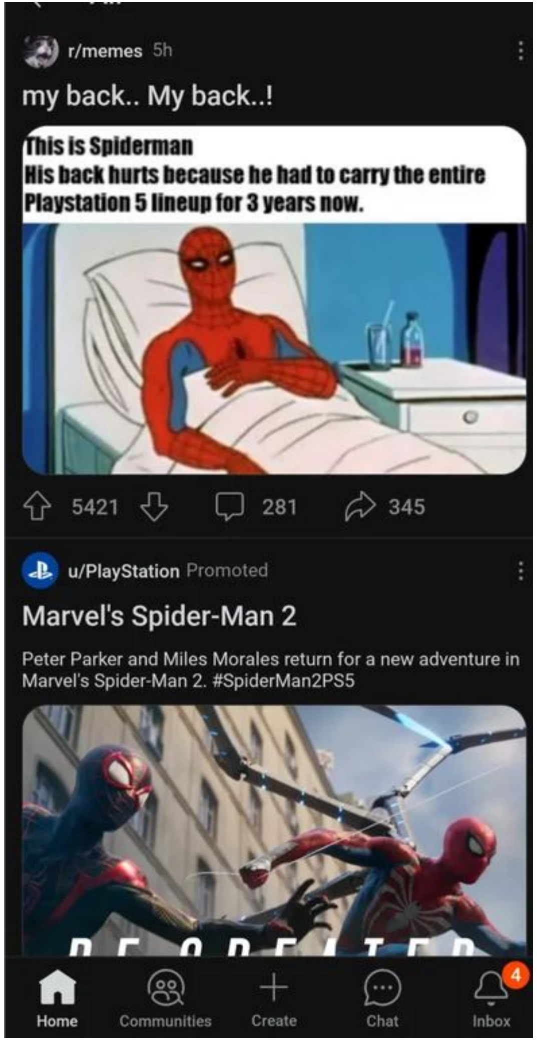
Analyze and data-mine funny internet images to gain insights on certain internet communities.

This is Spiderman
His back hurts because he had to carry the entire Playstation 5 lineup for 3 years now.

How to watch Breaking Bad :
Breaking Bad
The Pirate Bay
Suivre
Every Breaking Bad Frame In Order
@breaking_frames
made with mematic

Justification

- Memes often involve text-in-picture combined with a secondary image
- Textual information can provide insight for advertisements
  - What is being talked about
  - What is the sentiment
- The images are often similar and can provide context for the textual information
  - Comparison between two items
  - Sarcasm



r/memes 5h

my back.. My back..!

This is Spiderman
His back hurts because he had to carry the entire Playstation 5 lineup for 3 years now.

5421 281 345

u/PlayStation Promoted

Marvel's Spider-Man 2

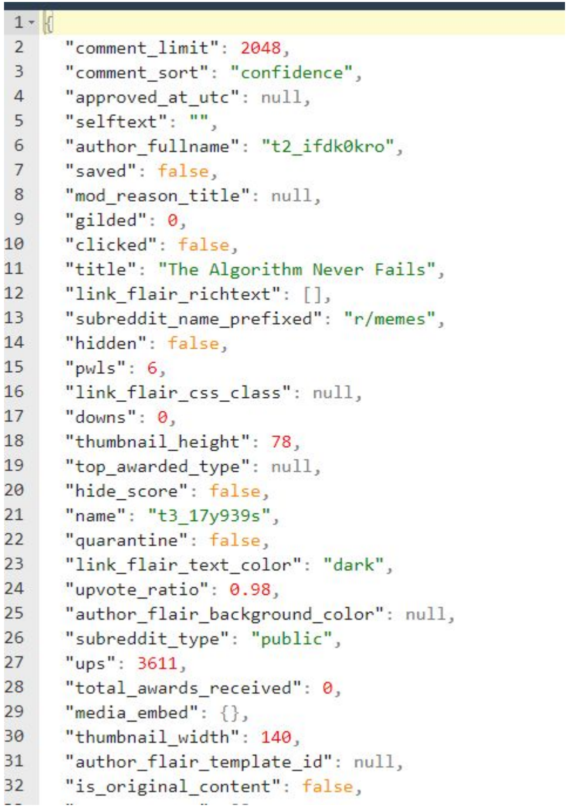
Peter Parker and Miles Morales return for a new adventure in Marvel's Spider-Man 2. #SpiderMan2PS5

Data sources

- Reddit
- Imgur
- ~~Ifunny~~ -> Tumblr
  - The reverse engineered iFunny API turned out to be outdated
  - We chose Tumblr as a replacement, which does have an official API
- ~~KnowYourMeme~~

Reddit

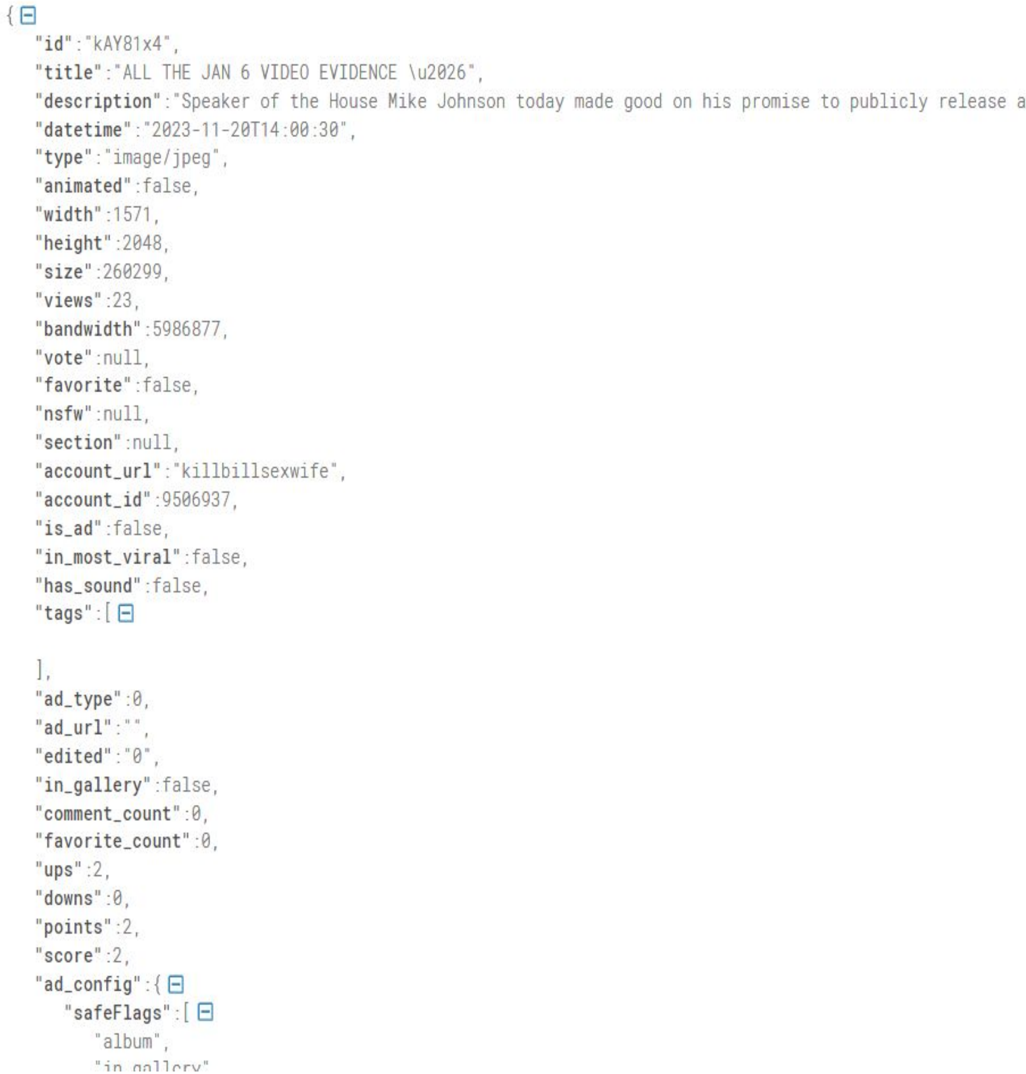
- We focused on reddit.com/r/memes
- Dedicated Python wrapper (PRAW)
- Hot, Top, New
- 100 requests per minute (managed by the wrapper)
    - We assumed that fetching info about one post is one request, but that might not be the case
- Around 1 new post/minute on /r/memes
    - Will likely expand to other, more active communities.



```
{
  "comment_limit": 2048,
  "comment_sort": "confidence",
  "approved_at_utc": null,
  "selftext": "",
  "author_fullname": "t2_ifdk0kro",
  "saved": false,
  "mod_reason_title": null,
  "gilded": 0,
  "clicked": false,
  "title": "The Algorithm Never Fails",
  "link_flair_richtext": [],
  "subreddit_name_prefixed": "r/memes",
  "hidden": false,
  "pwls": 6,
  "link_flair_css_class": null,
  "downs": 0,
  "thumbnail_height": 78,
  "top_awarded_type": null,
  "hide_score": false,
  "name": "t3_17y939s",
  "quarantine": false,
  "link_flair_text_color": "dark",
  "upvote_ratio": 0.98,
  "author_flair_background_color": null,
  "subreddit_type": "public",
  "ups": 3611,
  "total_awards_received": 0,
  "media_embed": {},
  "thumbnail_width": 140,
  "author_flair_template_id": null,
  "is_original_content": false,
```

Imgur

- Imgur-python library
- Hot, Top, New
- 12 500 requests per day
- Slight delay
- Around 1-2 new posts/minute
- Returns images or galleries
- Lots of additional metadata

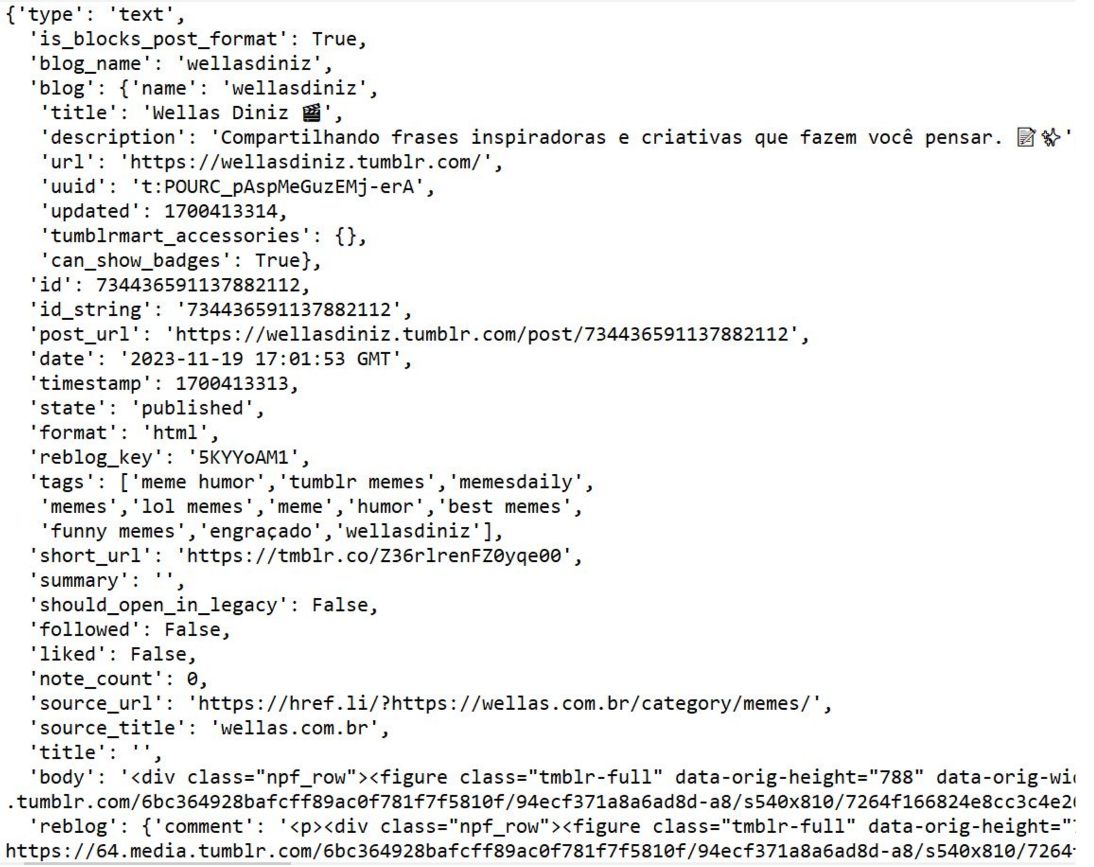


```
{⊟
   "id":"kAY81x4",
   "title":"ALL THE JAN 6 VIDEO EVIDENCE \u2026",
   "description":"Speaker of the House Mike Johnson today made good on his promise to publicly release a
   "datetime":"2023-11-20T14:00:30",
   "type":"image/jpeg",
   "animated":false,
   "width":1571,
   "height":2048,
   "size":260299,
   "views":23,
   "bandwidth":5986877,
   "vote":null,
   "favorite":false,
   "nsfw":null,
   "section":null,
   "account_url":"killbillsexwife",
   "account_id":9506937,
   "is_ad":false,
   "in_most_viral":false,
   "has_sound":false,
   "tags":[⊟

   ],
   "ad_type":0,
   "ad_url":"",
   "edited":"0",
   "in_gallery":false,
   "comment_count":0,
   "favorite_count":0,
   "ups":2,
   "downs":0,
   "points":2,
   "score":2,
   "ad_config":{⊟
      "safeFlags":[⊟
         "album",
         "in_gallery"
```

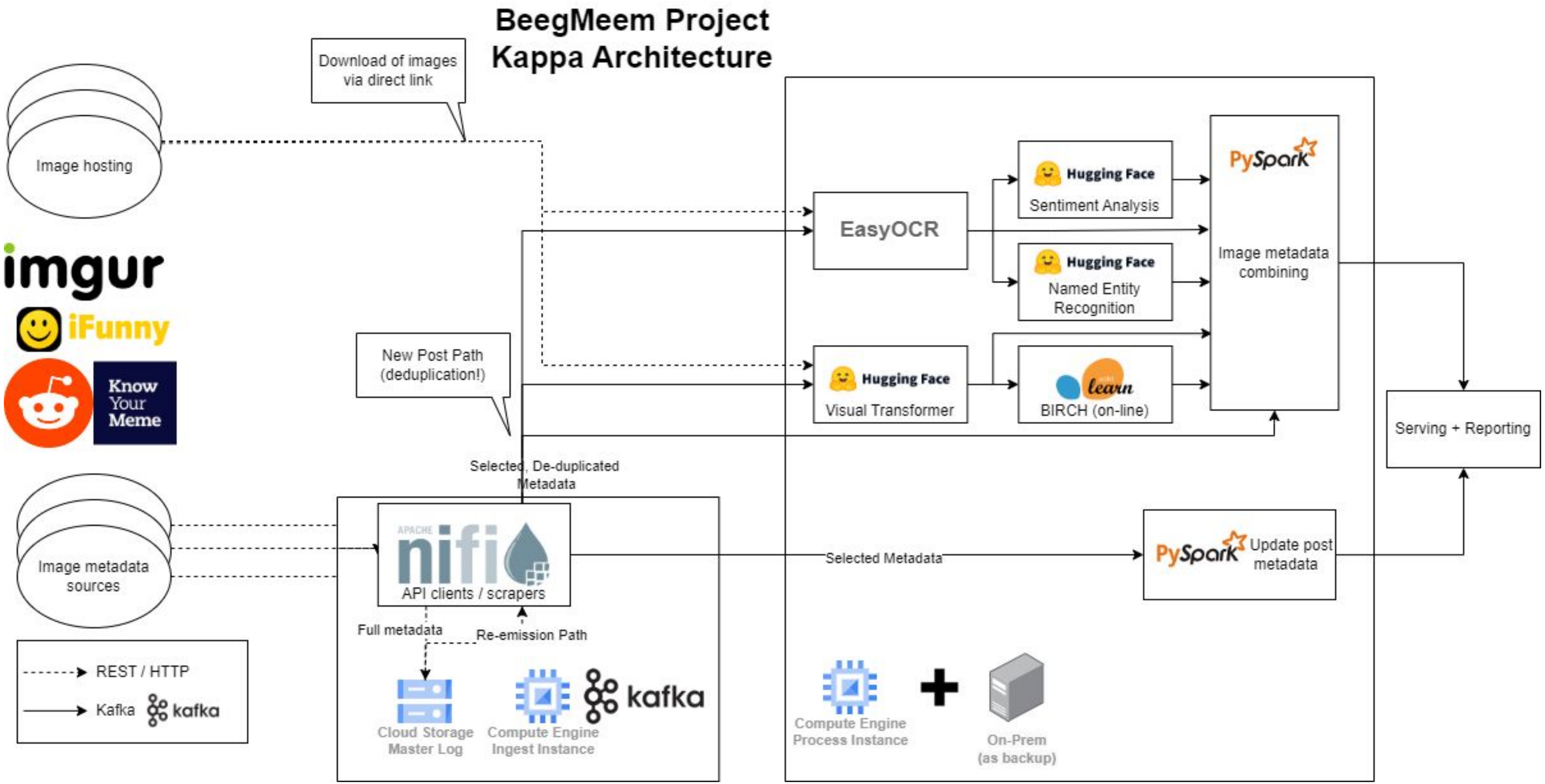
Tumblr

- pytumblr library
- latest memes in tags
- 1000 requests per hour
- 5000 requests per day
- 1 post/ 5-7 min in #memes
- Returns json with image urls
- Images of multiple resolutions
- Total number of interactions
- No number of views
- Author = blog name
- A lot of useless metadata



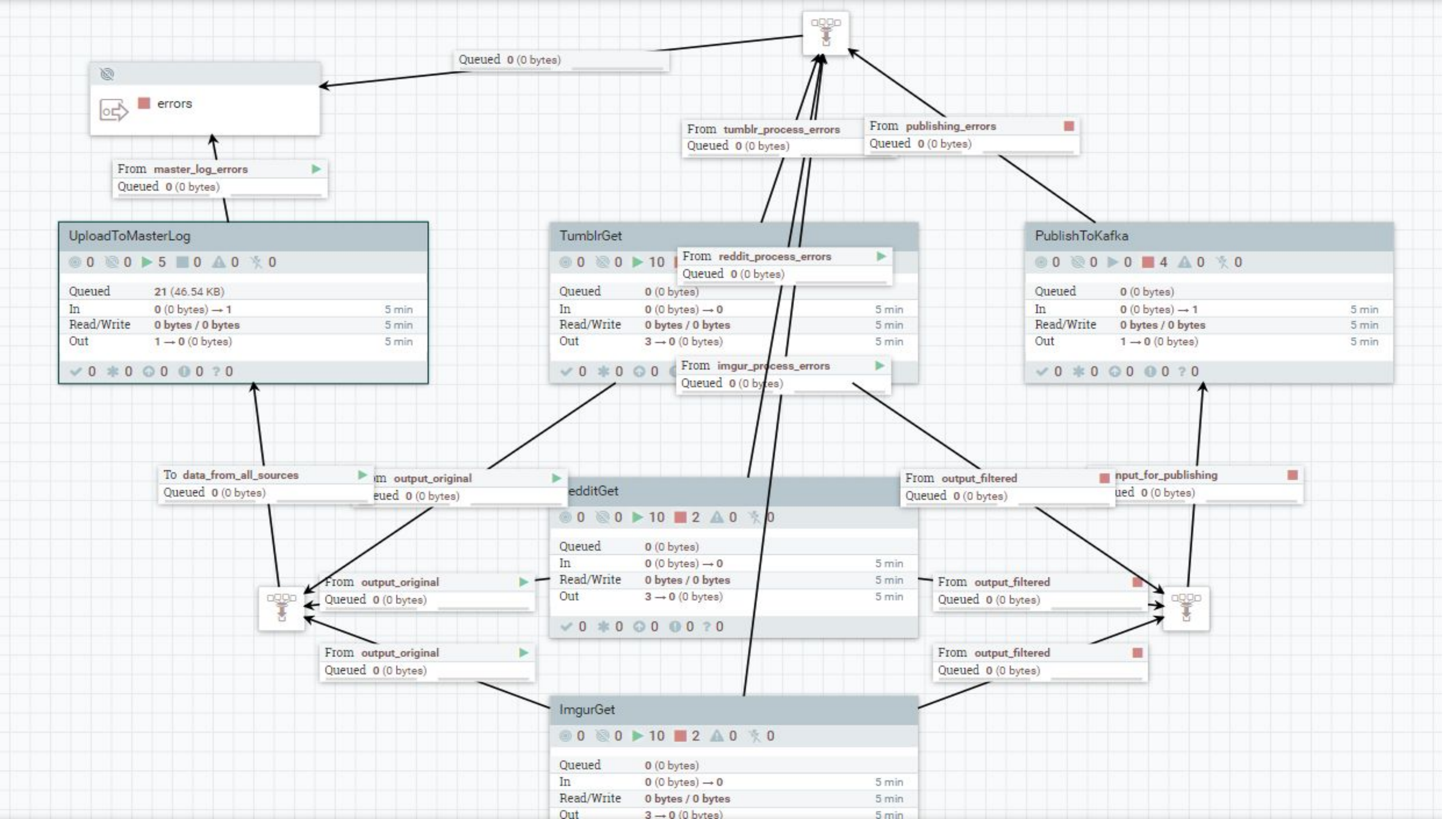
```
{'type': 'text',
  'is_blocks_post_format': True,
  'blog_name': 'wellasdiniz',
  'blog': {'name': 'wellasdiniz',
   'title': 'Wellas Diniz 🎬',
   'description': 'Compartilhando frases inspiradoras e criativas que fazem você pensar. 📝✨'
   'url': 'https://wellasdiniz.tumblr.com/',
   'uuid': 't:POURC_pAspMeGuzEMj-erA',
   'updated': 1700413314,
   'tumblrmart_accessories': {},
   'can_show_badges': True},
  'id': 734436591137882112,
  'id_string': '734436591137882112',
  'post_url': 'https://wellasdiniz.tumblr.com/post/734436591137882112',
  'date': '2023-11-19 17:01:53 GMT',
  'timestamp': 1700413313,
  'state': 'published',
  'format': 'html',
  'reblog_key': '5KYYoAM1',
  'tags': ['meme humor','tumblr memes','memesdaily',
   'memes','lol memes','meme','humor','best memes',
   'funny memes','engraçado','wellasdiniz'],
  'short_url': 'https://tmblr.co/Z36rlrenFZ0yqe00',
  'summary': '',
  'should_open_in_legacy': False,
  'followed': False,
  'liked': False,
  'note_count': 0,
  'source_url': 'https://href.li/?https://wellas.com.br/category/memes/',
  'source_title': 'wellas.com.br',
  'title': '',
  'body': '<div class="npf_row"><figure class="tmblr-full" data-orig-height="788" data-orig-wi
.tumblr.com/6bc364928bafcff89ac0f781f7f5810f/94ecf371a8a6ad8d-a8/s540x810/7264f166824e8cc3c4e2
  'reblog': {'comment': '<p><div class="npf_row"><figure class="tmblr-full" data-orig-height="
https://64.media.tumblr.com/6bc364928bafcff89ac0f781f7f5810f/94ecf371a8a6ad8d-a8/s540x810/7264
```

Architecture



Ingestion

- For data acquisition and preliminary transformation we use Apache Nifi
- Separate flows for each data source
- Data transformation, unification
- Writing to master log
- Publishing to Kafka



Queued 0 (0 bytes)

errors

From master_log_errors
Queued 0 (0 bytes)

From tumblr_process_errors
Queued 0 (0 bytes)

From publishing_errors
Queued 0 (0 bytes)

UploadToMasterLog
0 0 5 0 0 0
Queued 21 (46.54 KB)
In 0 (0 bytes) → 1 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 1 → 0 (0 bytes) 5 min
0 0 0 0 0

TumblrGet
0 0 10
Queued 0 (0 bytes)
In 0 (0 bytes) → 0 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 3 → 0 (0 bytes) 5 min
0 0 0

From reddit_process_errors
Queued 0 (0 bytes)

From imgur_process_errors
Queued 0 (0 bytes)

PublishToKafka
0 0 0 4 0 0
Queued 0 (0 bytes)
In 0 (0 bytes) → 1 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 1 → 0 (0 bytes) 5 min
0 0 0 0 0

To data_from_all_sources
Queued 0 (0 bytes)

m output_original
eued 0 (0 bytes)

edditGet
0 0 10 2 0 0
Queued 0 (0 bytes)
In 0 (0 bytes) → 0 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 3 → 0 (0 bytes) 5 min
0 0 0 0 0

From output_filtered
Queued 0 (0 bytes)

nput_for_publishing
ued 0 (0 bytes)

From output_original
Queued 0 (0 bytes)

From output_filtered
Queued 0 (0 bytes)

From output_original
Queued 0 (0 bytes)

From output_filtered
Queued 0 (0 bytes)

ImgurGet
0 0 10 2 0 0
Queued 0 (0 bytes)
In 0 (0 bytes) → 0 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 3 → 0 (0 bytes) 5 min

Unified Data fields

- Global ID
- Author of the post
- Time the post was created (UTC)
- Text description of the post
- Post score
- URL to the post image
- Image source (Imgur, Reddit, Tumblr)

Transport

- We experimented with different solutions: Pub/Sub, Pub/Sub lite & Kafka
  - Pub/Sub has no connector to Spark
  - The Pub/Sub lite connector didn’t work
  - We spent too much time on something that was supposed to save us time
- We eventually just set up Apache Kafka.
- It will be used to glue different parts of the system together.

Processing

- Various data-mining models operating on image data and data extracted from images
    - Visual Transformer ->
        - On-line Clustering
    - Optical Character Recognition ->
        - Sentiment Analysis
        - Named Entity Recognition
- Metadata processing for new posts
- Metadata processing for post updates

ML Models

- Optical Character Recognition
  - recognizes text inside a meme
  - EasyOCR python module
- Sentiment Analysis
  - identifies the polarity of the meme text
  - pretrained HuggingFace ‘sentiment-analysis’ pipeline
- Named Entity Recognition
  - extract named entities from the meme text
  - HuggingFace bert-base-ner model
- Visual Transformer
  - pretrained HuggingFace model
- On-line Clustering
  - BIRCH algorithm, scikit-learn implementation

Serving + Reporting

MongoDB:

- real time querying
- flexible schema design
- uses JSON
- integration with Apache Spark

Data will be partitioned by day

Report will be in PowerBI Desktop, which will connect via ODBC data source

Conclusion

- Goal: Analyze and data-mine funny internet images to gain insights on certain internet communities
- Use Kappa architecture for application
- Sources of memes: Reddit, Imgur, Tumblr (tested and working)
- NiFi and Apache Kafka
- 4 ML models to add insights into data
- MongoDB
- PowerBI Desktop reporting