Beeg Meme Project 2nd Deliverable Tests

Detailed requirements for testing:

- *Functional tests for data acquisition and pre-processing modules*
- *Quality tests for data transformation*
- *A tabular format for test cases including:*
    - *Test objective*
    - *Test steps*
    - *Expected result*
    - *Actual result (e.g. screenshot, log, summary report)*

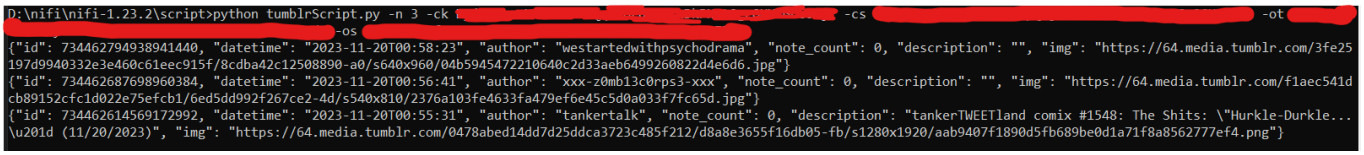
Test table

Test objective	Test steps	Expected Result	Actual Result
Test the script for acquiring Tumblr data	1.Data acquiring script was invoked with appropriate parameters (in our case we chose n=3 for clear printing)	Script prints out 3 json files with metadata about posted image (and link to the image itself)	As expected (proof below)
Test that the flow contains images from the source website	1.Start the data acquisition flow for Imgur 2.Collect some of the latest posts 3.Check the newest posts on the website	The data about posts in the flow is corresponding to the actual posts on the website	As expected (proof below)
Test data transformation (unification)	1.One flow file per data source is acquired 2.Flow files are transformed to be unified (have the same fields) 3. Unification result is checked on merged flowfile	The data in merged flow file have the same fields	As expected (proof below)
Test uploading data to the master log	1.All data source streams started acquiring data 2. When 100 files have been	New tar file has appeared in the cloud storage bucket	As expected (proof below)

	accumulated, they were then merged into a tar file 3. The tar file was sent to the gcp cloud storage bucket 4. New tar file was confirmed to have appeared in the cloud storage bucket		
Test publishing data records to Kafka topic	1.New Kafka topic named **test_topic** was created 2.Imgur data source stream was activated 3.After acquiring a few JSON files, they were then merged into a single file 4. KafkaPublisher process published all the JSON files inside the merged file	JSON files are visible in the Kafka topic	As expected (proof below)

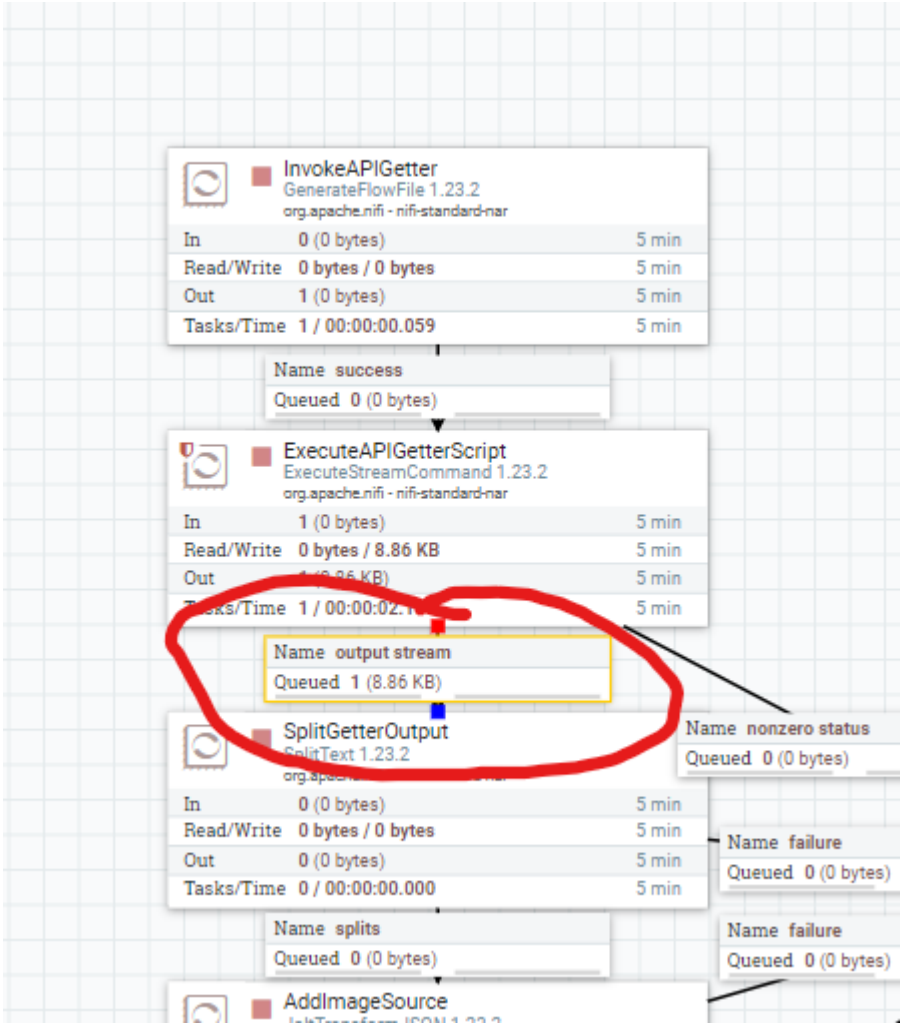
Tumblr script test

Here we can see the script invocation result. Sensitive data have been censored.

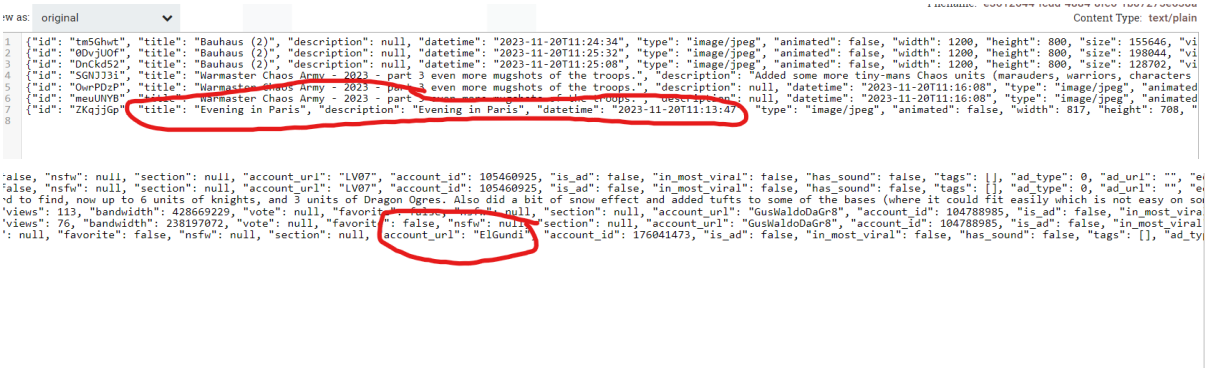


Imgur flow test

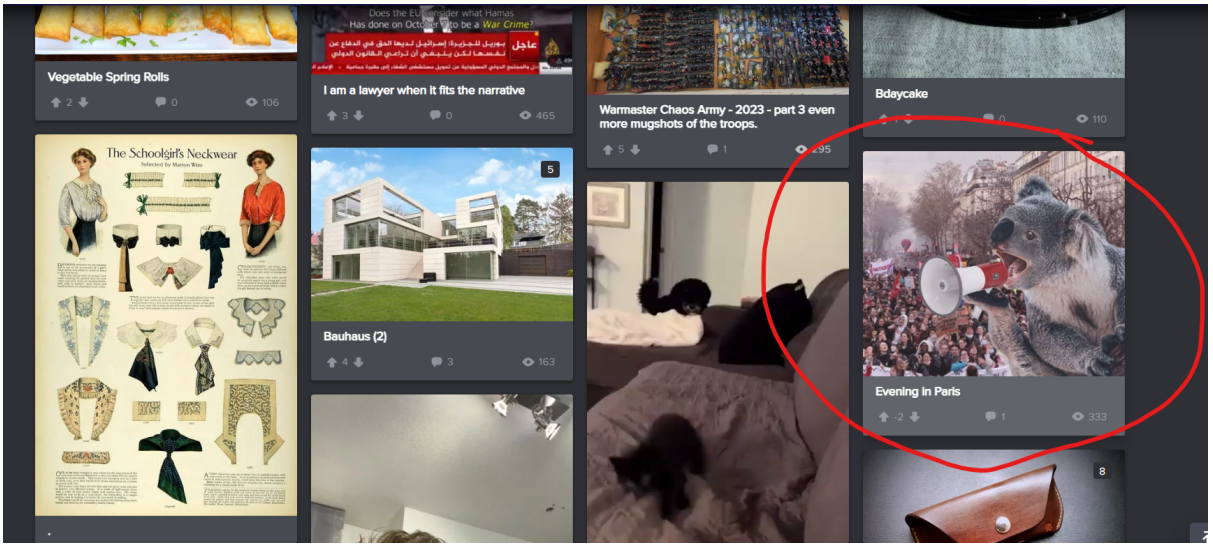
First we start the data acquisition flow for imgur. As we can see we have collected an output.

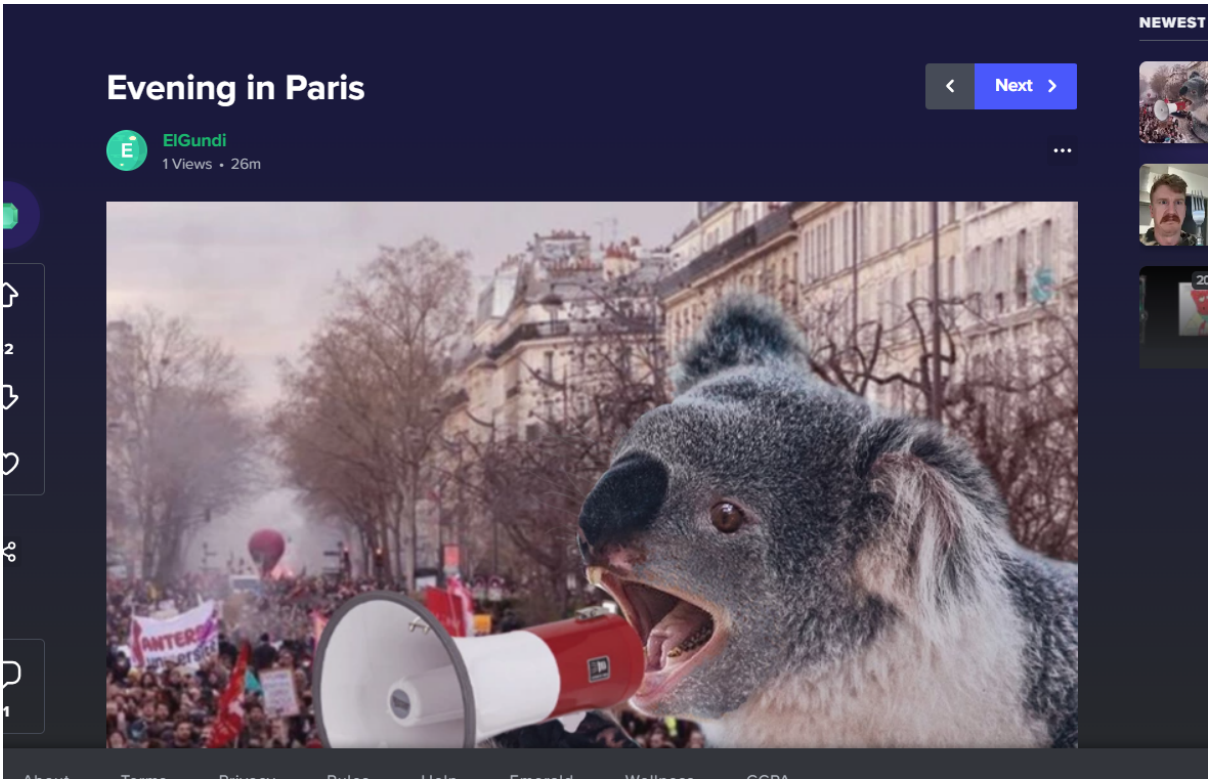


Let now us inspect the collected flowfile. We will fous on the last post (“Evening in Paris”), as it contains only 1 image.



We can now check the corresponding post on the website.





As we can see, the data is consistent with the post.

Data transformation (unification) test

First we acquire a single flowfile (json object) from each data source. Acquired objects:
First for Imgur:

```
{
  "id" : "CTtIanR",
  "title" : "My last sessions as a player in a nutshell",
  "description" : null,
  "datetime" : "2023-11-20T12:30:18",
  "type" : "image/jpeg",
  "animated" : false,
  "width" : 600,
  "height" : 600,
  "size" : 30256,
  "views" : 1,
  "bandwidth" : 30256,
  "vote" : null,
  "favorite" : false,
  "nsfw" : null,
  "section" : null,
  "account_url" : "EmpeRohr",
  "account_id" : 169264549,
  "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" : 1,
  "ups" : 1,
  "downs" : 0,
  "points" : 1,
  "score" : 1,
  "ad_config" : {
    "safeFlags" : [ "album", "in_gallery", "gallery" ],
    "highRiskFlags" : [ ],
    "unsafeFlags" : [ "sixth_mod_unsafe", "under_10", "updated_date" ],
    "wallUnsafeFlags" : [ ],
    "showsAds" : false,
    "showAdLevel" : 1,
    "safe_flags" : [ "album", "in_gallery", "gallery" ],
    "high_risk_flags" : [ ],
    "unsafe_flags" : [ "sixth_mod_unsafe", "under_10", "updated_date" ],
    "wall_unsafe_flags" : [ ],
```

```
  "nsfw" : null,
  "section" : null,
  "account_url" : "EmpeRohr",
  "account_id" : 169264549,
  "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" : 1,
  "ups" : 1,
  "downs" : 0,
  "points" : 1,
  "score" : 1,
  "ad_config" : {
    "safeFlags" : [ "album", "in_gallery", "gallery" ],
    "highRiskFlags" : [ ],
    "unsafeFlags" : [ "sixth_mod_unsafe", "under_10", "updated_date" ],
    "wallUnsafeFlags" : [ ],
    "showsAds" : false,
    "showAdLevel" : 1,
    "safe_flags" : [ "album", "in_gallery", "gallery" ],
    "high_risk_flags" : [ ],
    "unsafe_flags" : [ "sixth_mod_unsafe", "under_10", "updated_date" ],
    "wall_unsafe_flags" : [ ],
    "show_ads" : false,
    "show_ad_level" : 1,
    "nsfw_score" : 0.1
  },
  "is_album" : true,
  "topic" : null,
  "topic_id" : null,
  "global_id" : "CTtIanR-imgur",
  "created" : 1700483418,
  "created_utc" : 1700479818,
  "datetime_utc" : "2023-11-20T11:30:18",
  "url" : "https://i.imgur.com/CTtIanR.jpg",
  "imgSource" : "imgur"
}
```


Then Reddit:

```
{
  "comment_limit" : 2048,
  "comment_sort" : "confidence",
  "approved_at_utc" : null,
  "selftext" : "",
  "author_fullname" : "t2_cpmt47fww",
  "saved" : false,
  "mod_reason_title" : null,
  "gilded" : 0,
  "clicked" : false,
  "title" : "this is ridiculous",
  "link_flair_richtext" : [ ],
  "subreddit_name_prefixed" : "r/memes",
  "hidden" : false,
  "pwls" : 6,
  "link_flair_css_class" : null,
  "downs" : 0,
  "thumbnail_height" : 140,
  "top_awarded_type" : null,
  "hide_score" : false,
  "name" : "t3_17zikym",
  "quarantine" : false,
  "link_flair_text_color" : "dark",
  "upvote_ratio" : 0.83,
  "author_flair_background_color" : null,
  "subreddit_type" : "public",
  "ups" : 1178,
  "total_awards_received" : 0,
  "media_embed" : { },
  "thumbnail_width" : 140,
  "author_flair_template_id" : "5a9034e0-5b20-11ec-b095-fa0d11f15f9a",
  "is_original_content" : false,
  "user_reports" : [ ],
  "secure_media" : null,
  "is_reddit_media_domain" : true,
  "is_meta" : false,
  "category" : null,
  "secure_media_embed" : { },
  "link_flair_text" : null,
  "can_mod_post" : false,
  "score" : 1178,
  "approved_by" : null,
  "is_created_from_ads_ui" : false,
```

```
"link_flair_text" : null,
"can_mod_post" : false,
"score" : 1178,
"approved_by" : null,
"is_created_from_ads_ui" : false,
"author_premium" : true,
"thumbnail" : "https://b.thumbs.redditmedia.com/Fhu5xdrGvf040e5K0NK7jk5UTrDL1E_nzRk8HRUiZfQ.jpg",
"edited" : false,
"author_flair_css_class" : null,
"author_flair_richtext" : [ {
  "e" : "text",
  "t" : "Shitposter"
} ],
"gildings" : { },
"post_hint" : "image",
"content_categories" : null,
"is_self" : false,
"mod_note" : null,
"created" : 1.700463233E9,
"link_flair_type" : "text",
"wls" : 6,
"removed_by_category" : null,
"banned_by" : null,
"author_flair_type" : "richtext",
"domain" : "i.redd.it",
"allow_live_comments" : false,
"selftext_html" : null,
"likes" : null,
"suggested_sort" : null,
"banned_at_utc" : null,
"url_overridden_by_dest" : "https://i.redd.it/87fjeel8ag1c1.jpeg",
"view_count" : null,
"archived" : false,
"no_follow" : false,
"is_crosspostable" : false,
"pinned" : false,
"over_18" : false,
"preview" : {
  "images" : [ {
    "source" : {
      "url" : "https://preview.redd.it/87fjeel8ag1c1.jpeg?auto=webp&s=836a15c395cccd145eec088ce702c3f52a263973",
      "width" : 1400,
      "height" : 1752
    }
```

```
      "height" : 135
    }, {
      "url" : "https://preview.redd.it/87fjeel8ag1c1.jpeg?width=216&crop=smart&auto=webp&s=7e815bb2ebc3bacc9d23080a76900d8baa18f583",
      "width" : 216,
      "height" : 270
    }, {
      "url" : "https://preview.redd.it/87fjeel8ag1c1.jpeg?width=320&crop=smart&auto=webp&s=2d2ed960c689d08b566eb451f98b4d93197ba1fe",
      "width" : 320,
      "height" : 400
    }, {
      "url" : "https://preview.redd.it/87fjeel8ag1c1.jpeg?width=640&crop=smart&auto=webp&s=404be8a43866c32f56c13a801bfcdeaf8f638db8",
      "width" : 640,
      "height" : 800
    }, {
      "url" : "https://preview.redd.it/87fjeel8ag1c1.jpeg?width=960&crop=smart&auto=webp&s=0cc78775caeacede4d5b2af7cf59ca76e812ae6e",
      "width" : 960,
      "height" : 1201
    }, {
      "url" : "https://preview.redd.it/87fjeel8ag1c1.jpeg?width=1080&crop=smart&auto=webp&s=0d64121309b7f3c637607f3b37a063502ee5c925",
      "width" : 1080,
      "height" : 1351
    } ],
    "variants" : { },
    "id" : "vfUb76wOFd_zSjHhcjeyQRYhE8wC6dvvSAvyEMJNKzk"
  } ],
  "enabled" : true
},
"all_awardings" : [ ],
"awarders" : [ ],
"media_only" : false,
"can_gild" : false,
"spoiler" : false,
"locked" : false,
"author_flair_text" : "Shitposter",
"treatment_tags" : [ ],
"visited" : false,
"removed_by" : null,
"num_reports" : null,
"distinguished" : null,
"subreddit_id" : "t5_2qjpg",
"author_is_blocked" : false,
"mod_reason_by" : null,
"removal_reason" : null,
```


As we can see, both Imgur and Reddit have multiple additional fields in their sources. Finally Tumblr:


```
{
  "id" : 734510693241536512,
  "datetime" : "2023-11-20T12:39:43",
  "author" : "a--z--u--1",
  "note_count" : 0,
  "description" : "",
  "img" : "https://64.media.tumblr.com/db94f7e107f6f41da10eb23e05f6d03f/bacbccc6f7a6f99e-1b/s1280x1920/7fe68c712b72ba44907f48ddf7026f0c30b8eda5.jpg",
  "imgSource" : "tumblr",
  "global_id" : "734510693241536512-tumblr"
}
```

Having acquired all the flow files, they are then unified and merged. Here is the result of unification and merging:

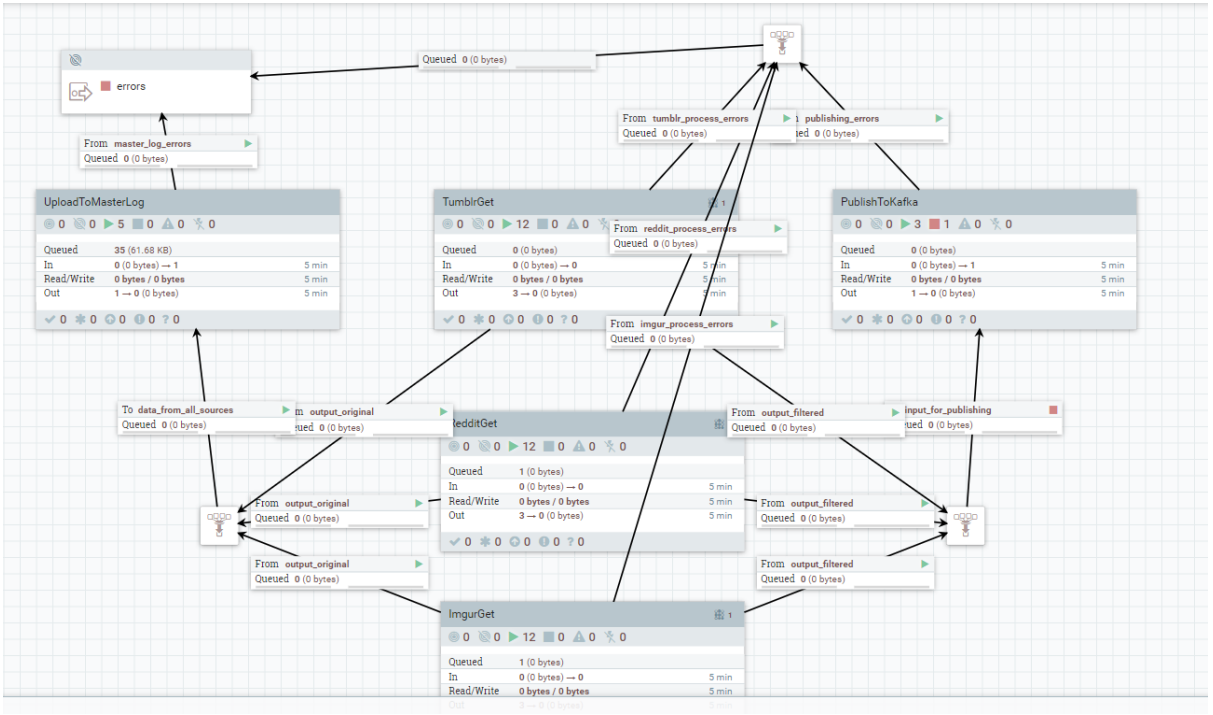

```
{"global_id":"CTtIanR-imgur","author":"EmpeRohr","created_time":"2023-11-20T11:30:18","desc":null,"score":1,"url":"https://i.imgur.com/CTtIanR.jpg","source":"imgur"}
{"global_id":"17zikym-reddit","author":"t2_cpmt47fww","created_time":"2023-11-20T06:53:53","desc":"","score":1178,"url":"https://i.redd.it/87fjeel8ag1c1.jpeg","source":"reddit"}
{"global_id":"734510693241536512-tumblr","author":"a--z--u--l","created_time":"2023-11-20T12:39:43","desc":"","score":0,"url":"https://64.media.tumblr.com/db94f7e107f6f41da10eb23e05f6d03f
```

```
':"https://i.imgur.com/CTtIanR.jpg","source":"imgur"}
},"url":"https://i.redd.it/87fjeel8ag1c1.jpeg","source":"reddit"}
:ore":0,"url":"https://64.media.tumblr.com/db94f7e107f6f41da10eb23e05f6d03f/bacbccc6f7a6f99e-1b/s1280x1920/7fe68c712b72ba44907f48ddf7026f0c30b8eda5.jpg","source":"tumblr"}
```

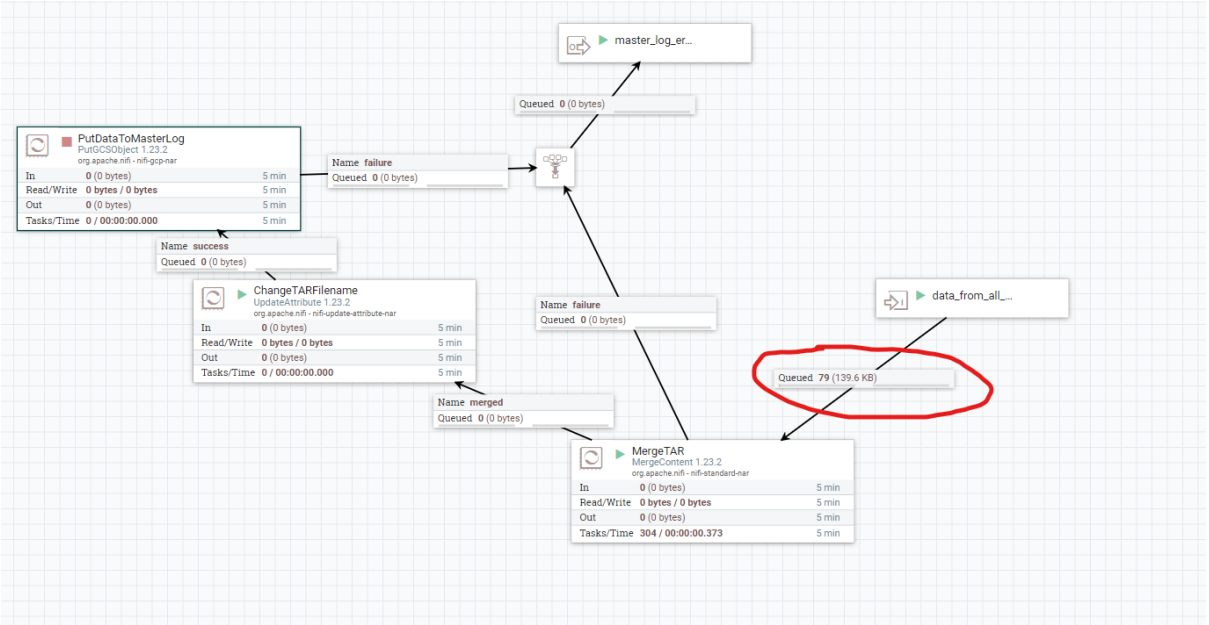

As we can see, now all the files have the same fields, with consistent names.

Master Log test

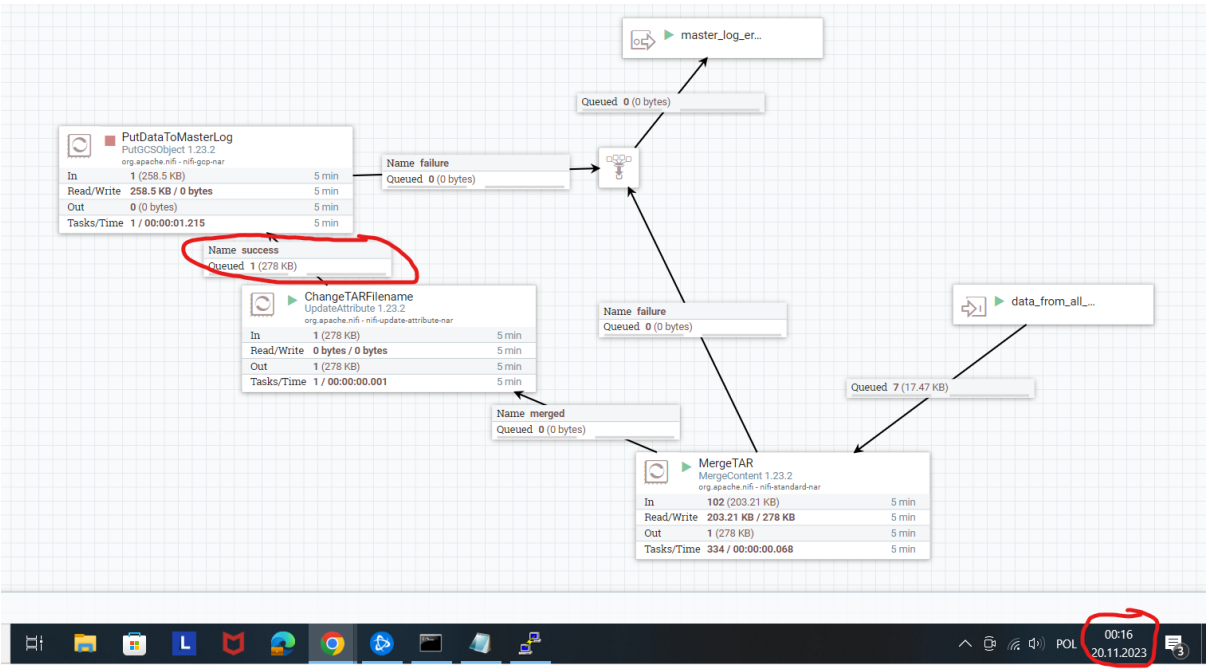
First we start all the data acquisition streams



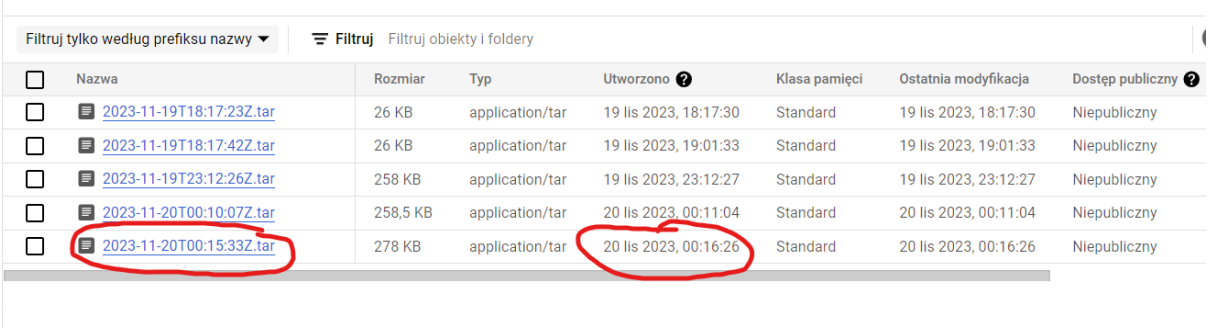
As we can see, data is being acquired, as it waits in queue to be merged (it requires at least 100 flowfiles)



After merging, the newly created flowflie is sent to be uploaded to cloud storage. Take notice of the time of sending.

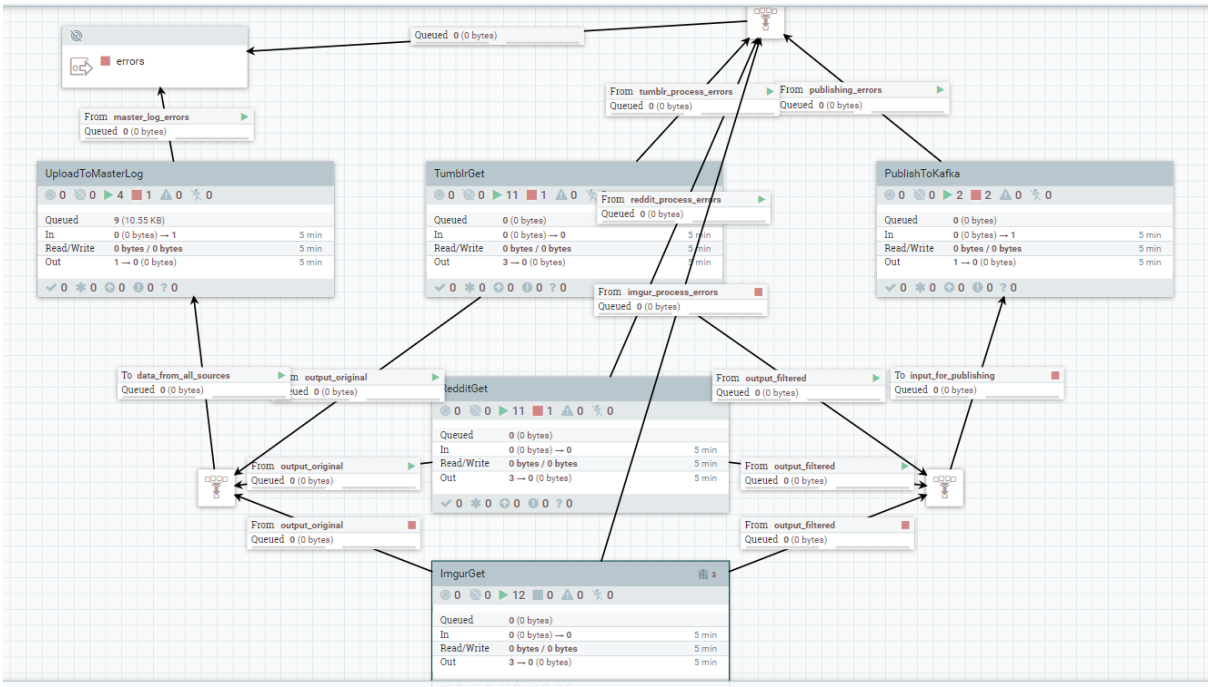


Finally, as we can see, new tar file has been uploaded to the master log. The time of creation is consistent with the time in which it was sent.

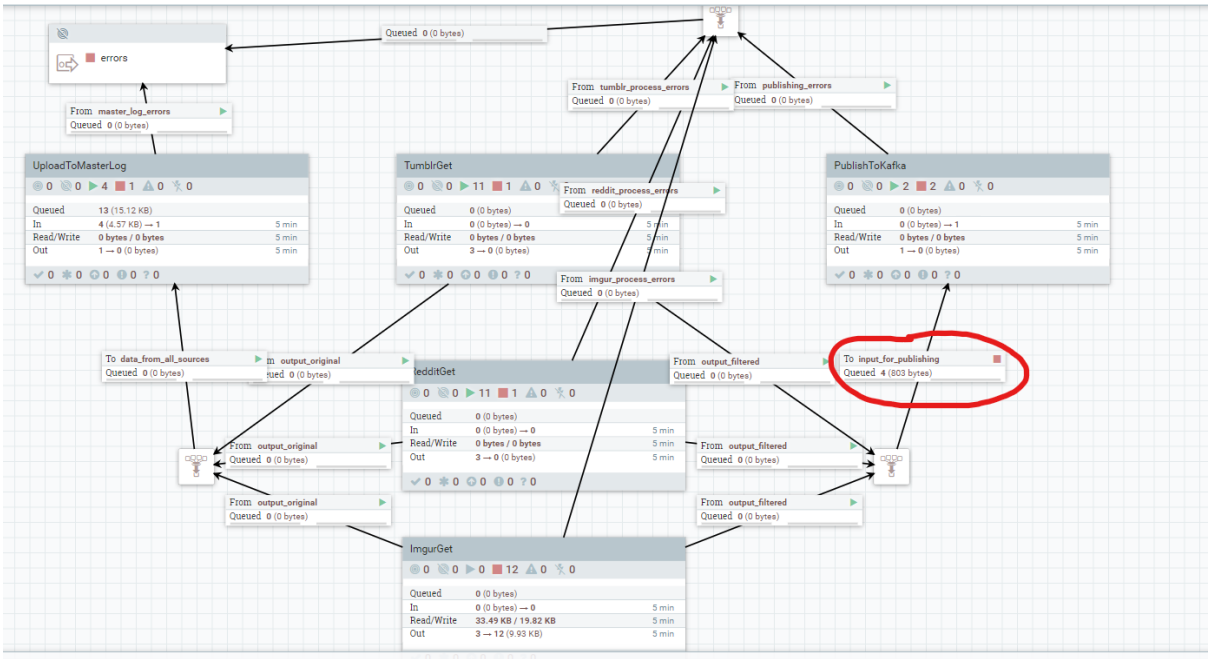


Kafka publishing test

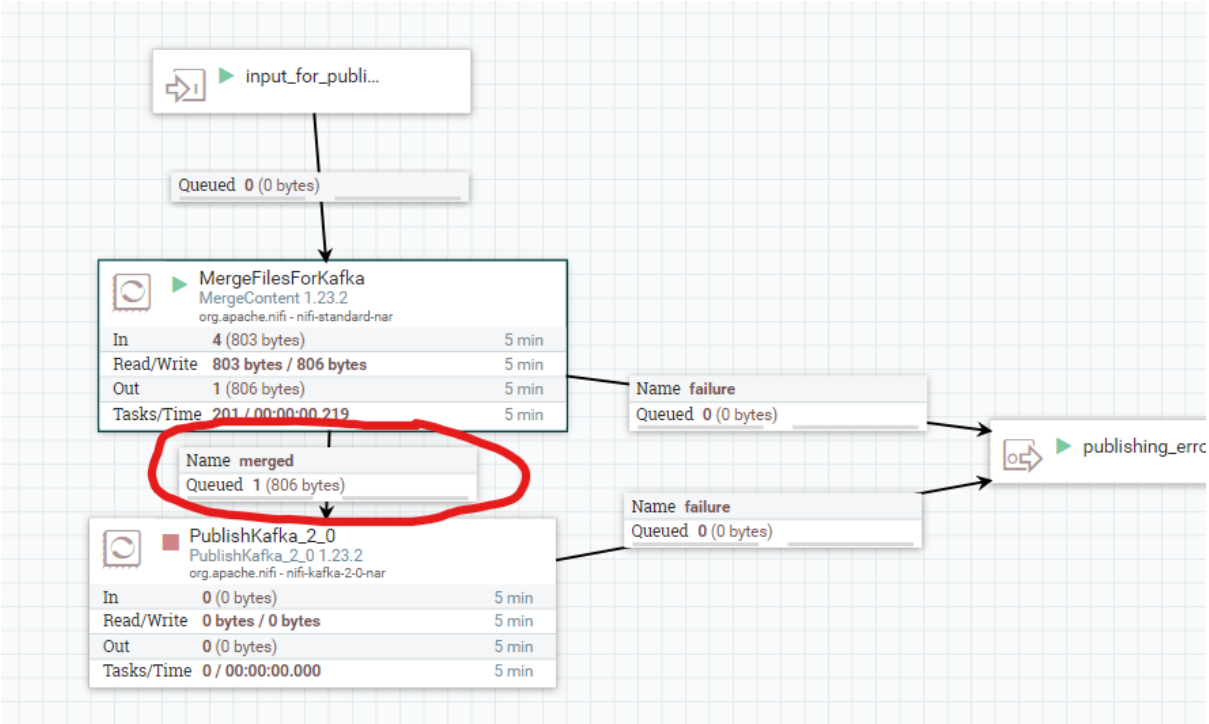
First, we start the data acquisition stream for Imgur.



As we can see, the files are being queued to be merged into one before being published.



Finally, the files have been merged.



Now we start the PublishKafka processor and publish all the records to the newly created kafka topic. We can compare the files as seen in nifi queue and on kafka topic – as we can see, those are the same files.
First nifi queue:


```
{"global_id":"CZwQnE4-imgur","author":"hollyrockB","created_time":"2023-11-19T23:02:36","desc":null,"score":8,"url":"https://i.imgur.com/CZwQnE4.png","source":"imgur"}
{"global_id":"Acs9ZiA-imgur","author":"hollyrockB","created_time":"2023-11-19T23:02:36","desc":null,"score":8,"url":"https://i.imgur.com/Acs9ZiA.png","source":"imgur"}
{"global_id":"2q6GHCO-imgur","author":"hollyrockB","created_time":"2023-11-19T23:02:36","desc":null,"score":8,"url":"https://i.imgur.com/2q6GHCO.jpg","source":"imgur"}
{"global_id":"oIAt2As-imgur","author":"MaiseyTolkien","created_time":"2023-11-19T22:56:47","desc":"I have never done this before this will be my first time !! Does anyone have any tips or
```


And then Kafka topic:


```
jakub_foltyn1217@nifi:/home/nifi/kafka_2.13-3.6.0$ sudo bin/kafka-console-consumer.sh --topic test_topic --from-beginning --bootstrap-server localhost:9092
{"global_id":"CZwQnE4-imgur","author":"hollyrockB","created_time":"2023-11-19T23:02:36","desc":null,"score":8,"url":"https://i.imgur.com/CZwQnE4.png","source":"imgur
"}
{"global_id":"Acs9ZiA-imgur","author":"hollyrockB","created_time":"2023-11-19T23:02:36","desc":null,"score":8,"url":"https://i.imgur.com/Acs9ZiA.png","source":"imgur
"}
{"global_id":"2q6GHCO-imgur","author":"hollyrockB","created_time":"2023-11-19T23:02:36","desc":null,"score":8,"url":"https://i.imgur.com/2q6GHCO.jpg","source":"imgur
"}
{"global_id":"oIAt2As-imgur","author":"MaiseyTolkien","created_time":"2023-11-19T22:56:47","desc":"I have never done this before this will be my first time !! Does a
nyone have any tips or helpful ideas how this all works ? ? ? Thanks","score":3,"url":"https://i.imgur.com/oIAt2As.png","source":"imgur"}
^CProcessed a total of 4 messages
```