# MESSY WEB, MEET CLEAN(ER) WEB: AN EXPERIMENT IN SOCIAL MEDIA ARCHIVING

Rachel Trent
Digital Services Manager
George Washington University Libraries, Special Collections

Archive-It Annual Partners Meeting
8/2/2016
Atlanta, GA

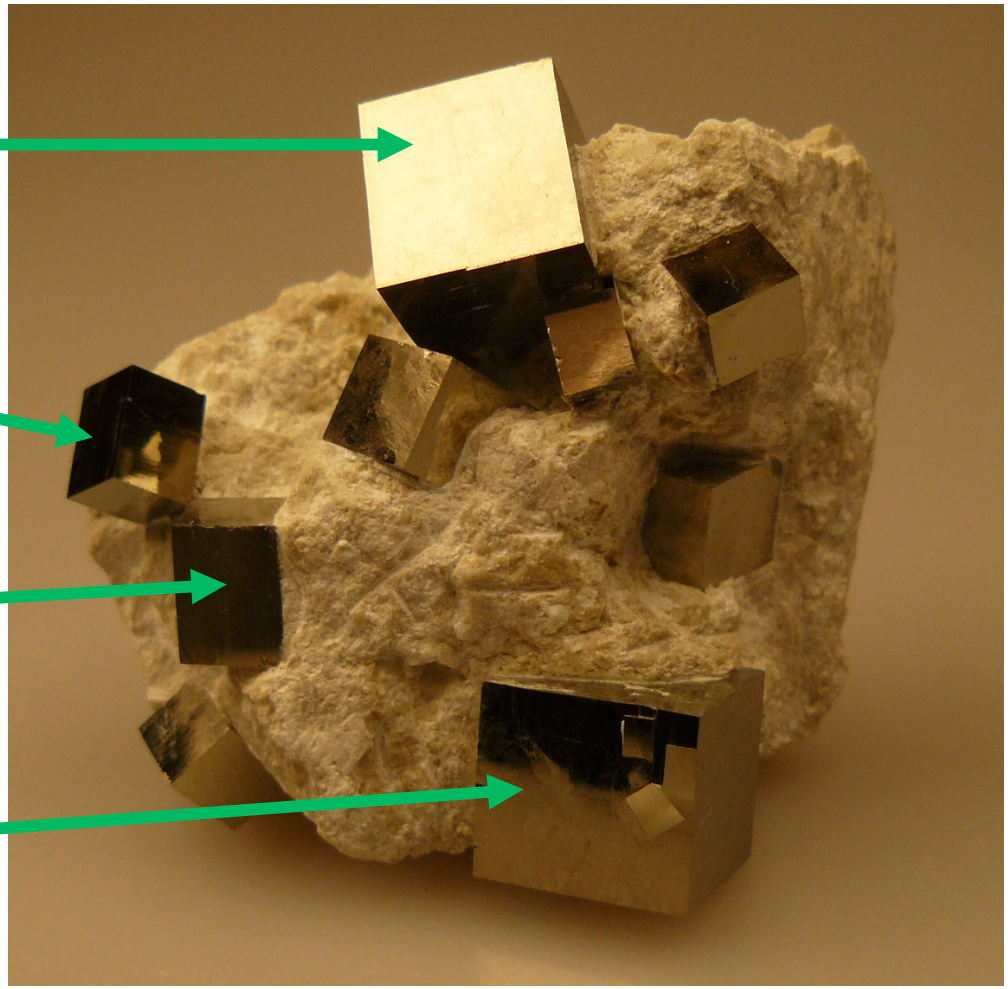
# THE WEB IS MESSY

TWITTER

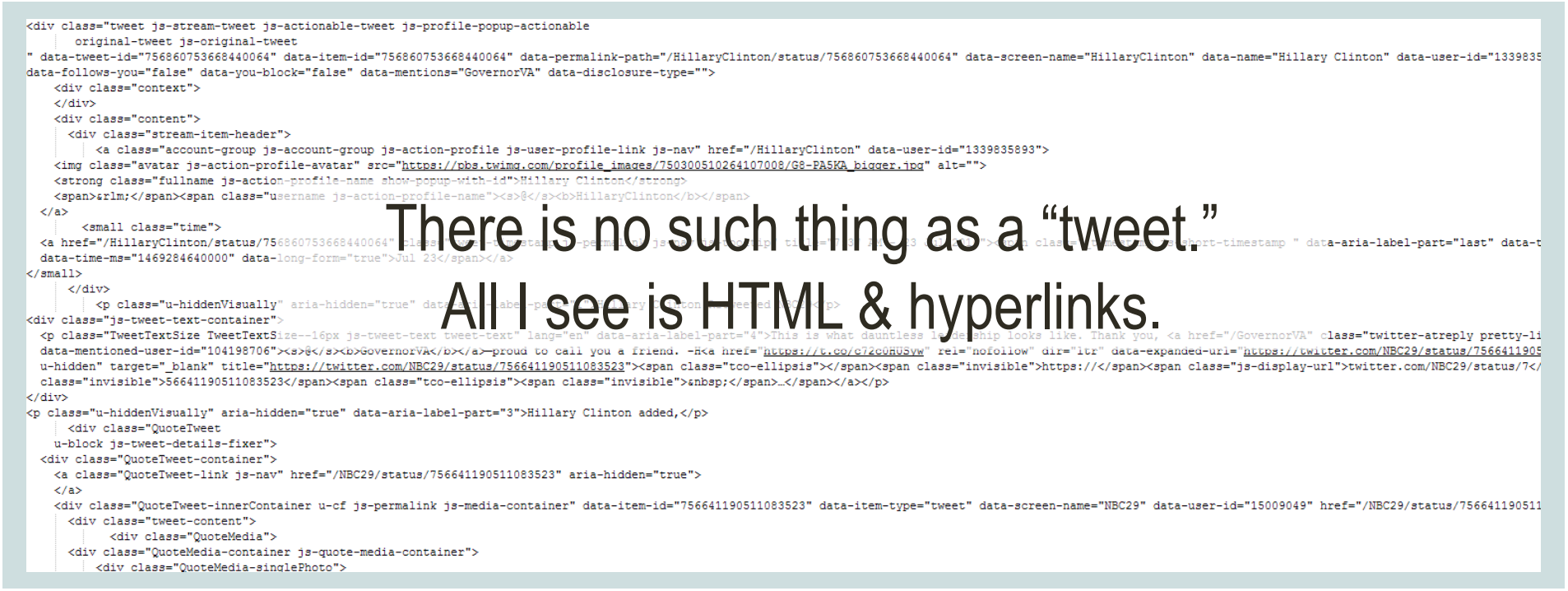
FACEBOOK

INSTAGRAM

WIKIPEDIA



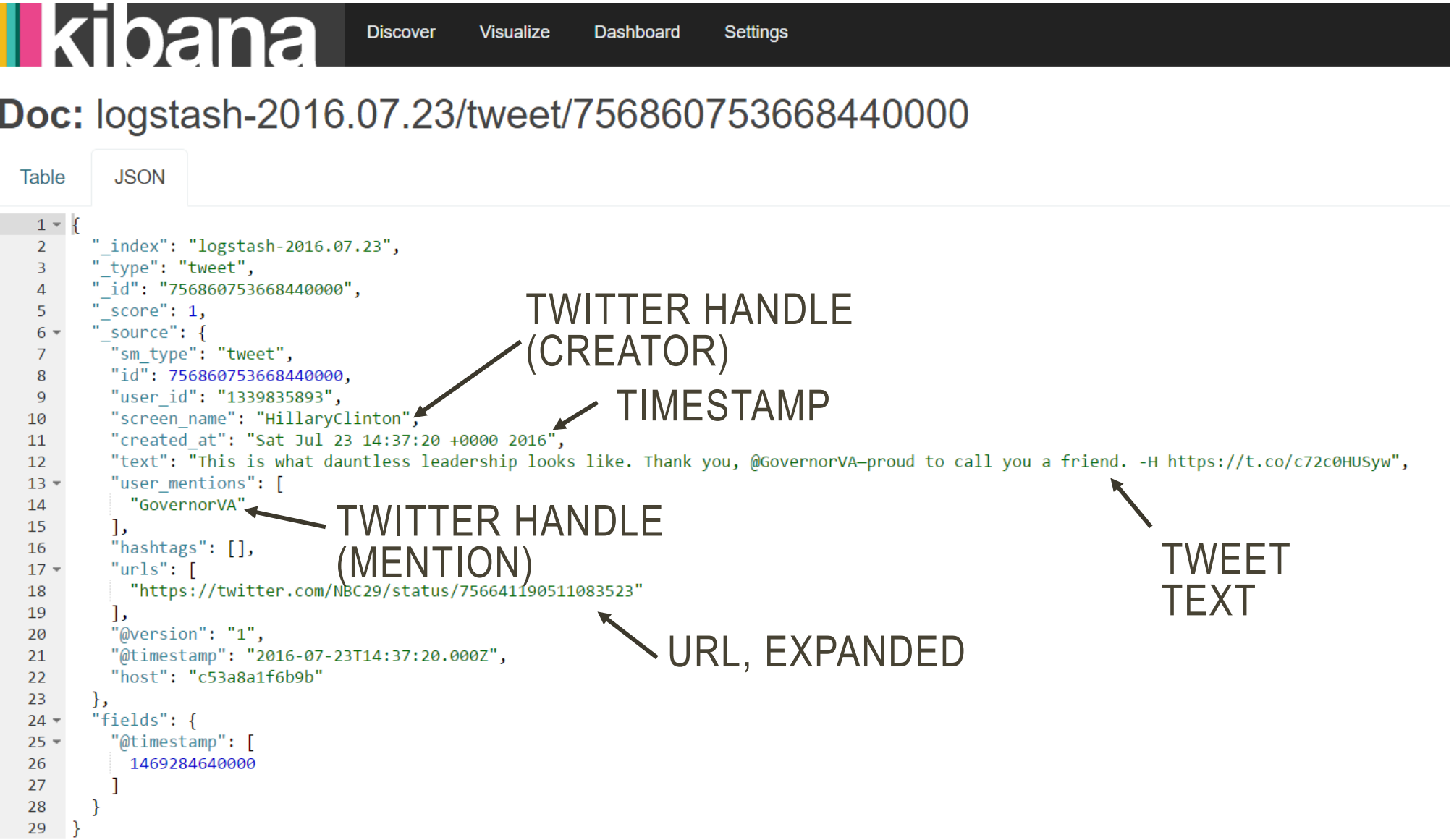

# WHAT A PERSON THINKS A TWEET IS




## WHAT A CRAWLER THINKS A TWEET IS

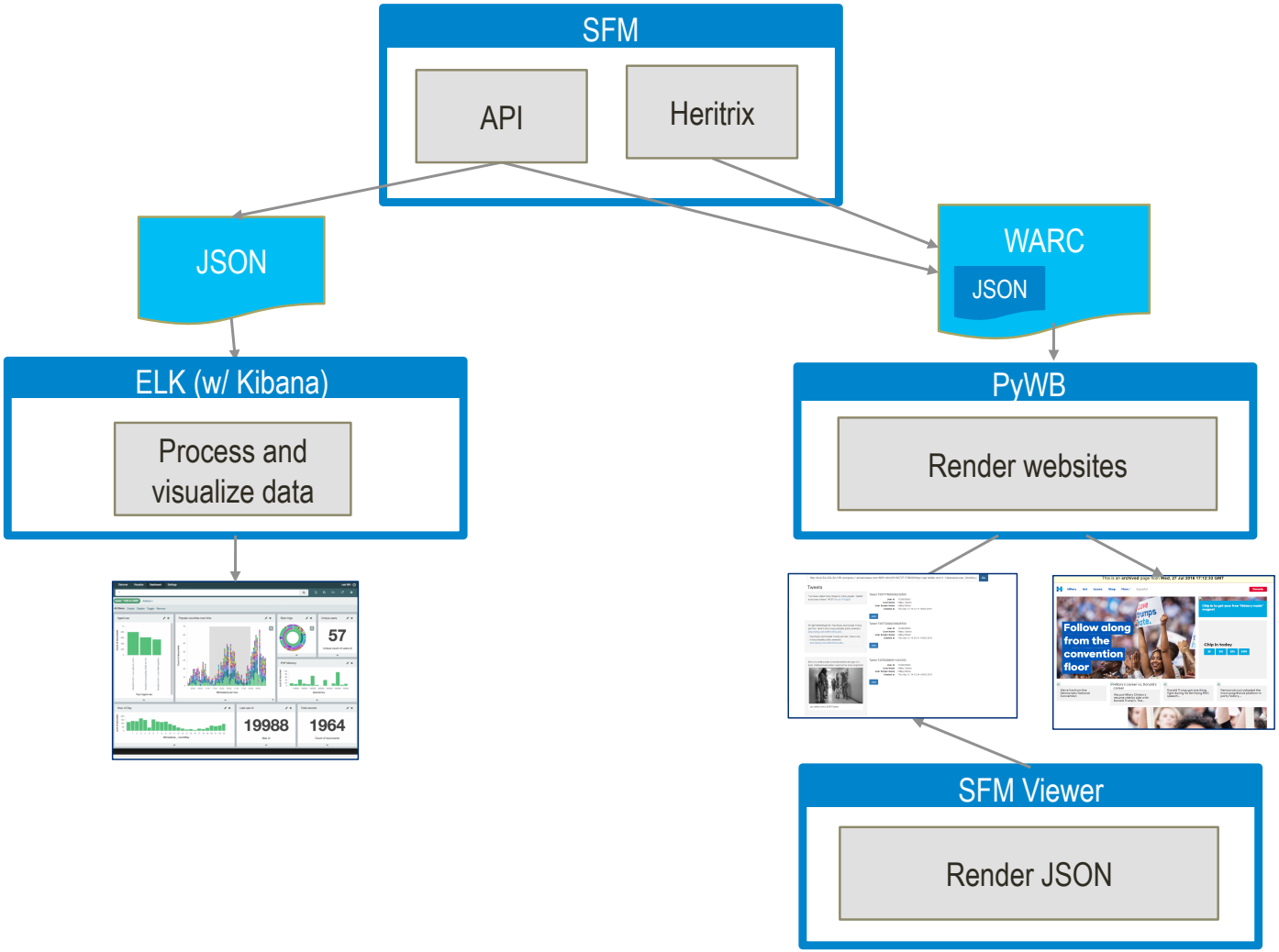
```
{
    "created_at": "Sat Jul 23 14:37:20 +0000 2016",
    "id": 756860753668440000,
    "id_str": "756860753668440064",
    "text": "This is what dauntless leadership looks like. Thank you, @GovernorVA—proud to
    call you a friend. -H https://t.co/c72c0HUSyw",
    "truncated": false,
    "entities": {
        "hashtags": [],
        "symbols": [],
        "user_mentions": [
            {
                "screen_name": "GovernorVA",
                "name": "Terry McAuliffe",
```


# WHAT AN APP THINKS A TWEET IS





```
{
  "_index": "logstash-2016.07.23",
  "_type": "tweet",
  "_id": "756860753668440000",
  "_score": 1,
  "_source": {
    "sm_type": "tweet",
    "id": 756860753668440000,
    "user_id": "1339835893",
    "screen_name": "HillaryClinton",
    "created_at": "Sat Jul 23 14:37:20 +0000 2016",
    "text": "This is what dauntless leadership looks like. Thank you, @GovernorVA—proud to call you a friend. -H https://t.co/c72c0HUSyw",
    "user_mentions": [
      "GovernorVA"
    ],
    "hashtags": [],
    "urls": [
      "https://twitter.com/NBC29/status/756641190511083523"
    ],
    "@version": "1",
    "@timestamp": "2016-07-23T14:37:20.000Z",
    "host": "c53a8a1f6b9b"
  },
  "fields": {
    "@timestamp": [
      1469284640000
    ]
  }
}
```


SFM
API
Heritrix
JSON
WARC
JSON
ELK (w/ Kibana)
Process and visualize data
PyWB
Render websites
57
19988
1964
Follow along from the convention floor
SFM Viewer
Render JSON

# DEMO

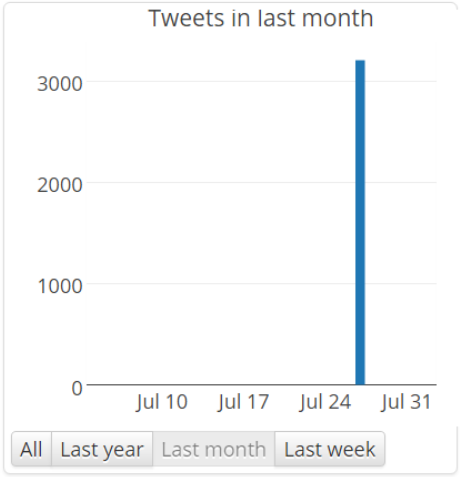
SOCIAL FEED MANAGER (SFM)
ELK

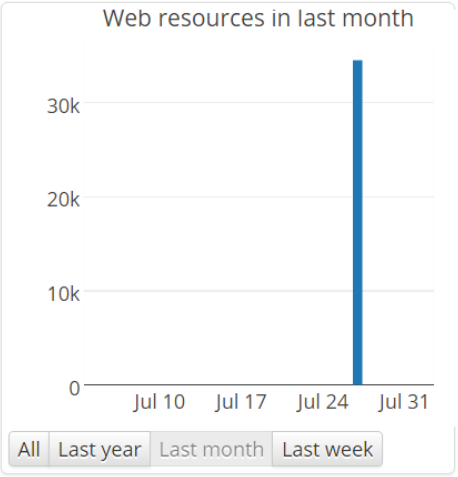
PYWB
SFM VIEWER

(screenshots if web fails us)

Social Feed Manager  Collection Sets  Credentials  Exports  Welcome, demo

# Demo - Hillary Clinton



Collection Sets / Demo - Hillary Clinton

# Demo - Hillary Clinton

Edit

Collection of campaign content for Hillary Clinton's campaign for United States President, 2016

**Group:** demo

**Stats:**

- tweets: 3210
- web resources: 34521

**Id:** ced319ce6c1d414592a213ae77feab87

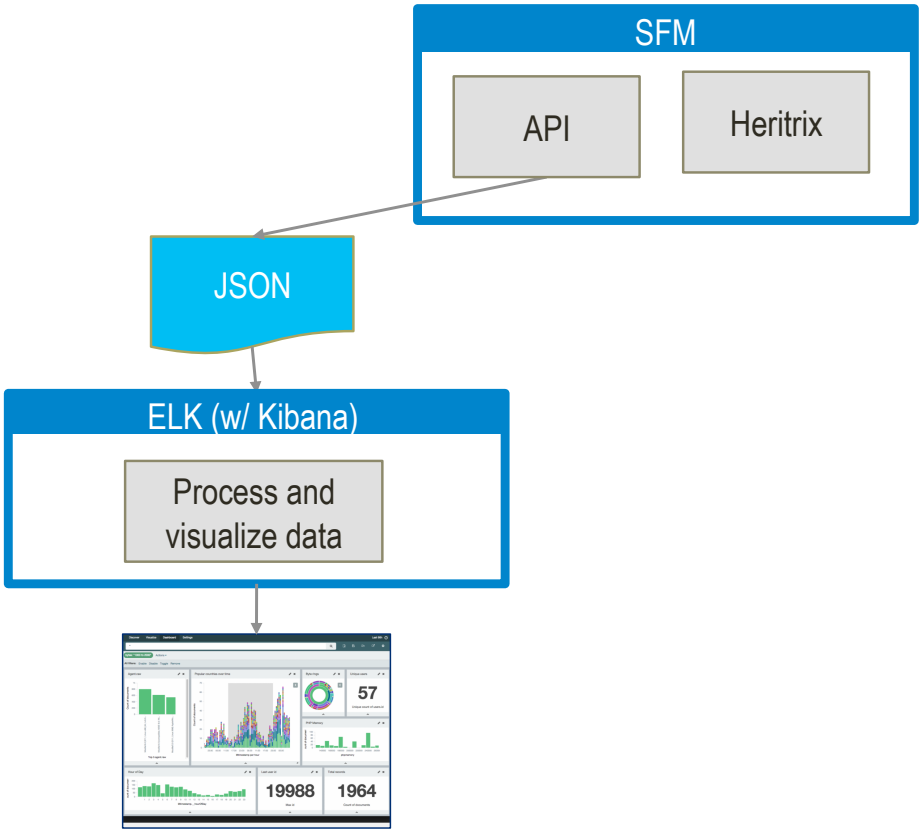
**Created:** July 27, 2016, 12:52 p.m.

## Collections

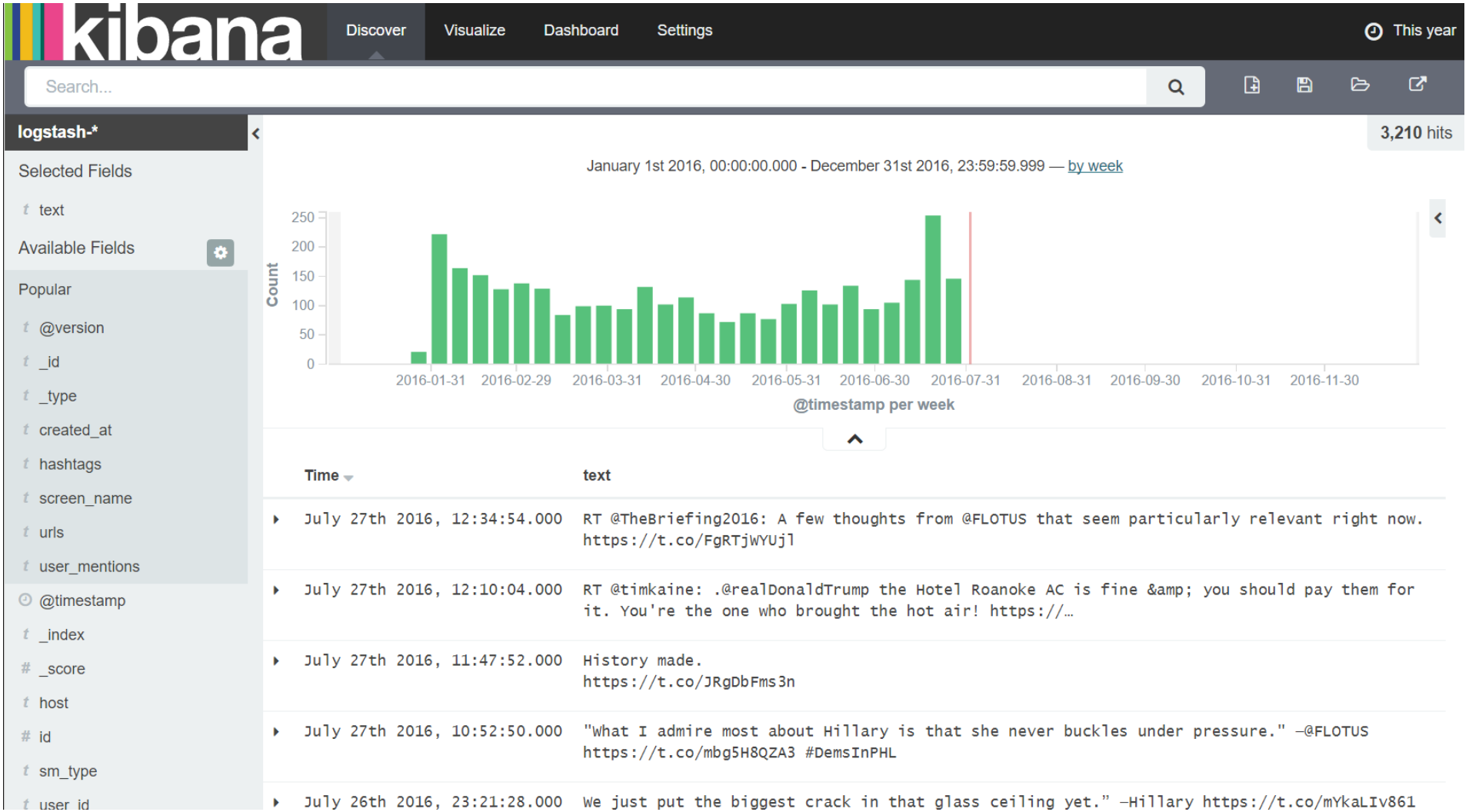
| Name | Harvest type | Seeds | On/off |
|---|---|---|---|
| @HillaryClinton | Twitter user timeline | 1 seed | Off |

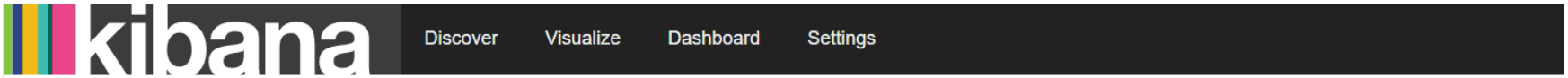
Add Collection

Change log




(screenshots if web fails us)

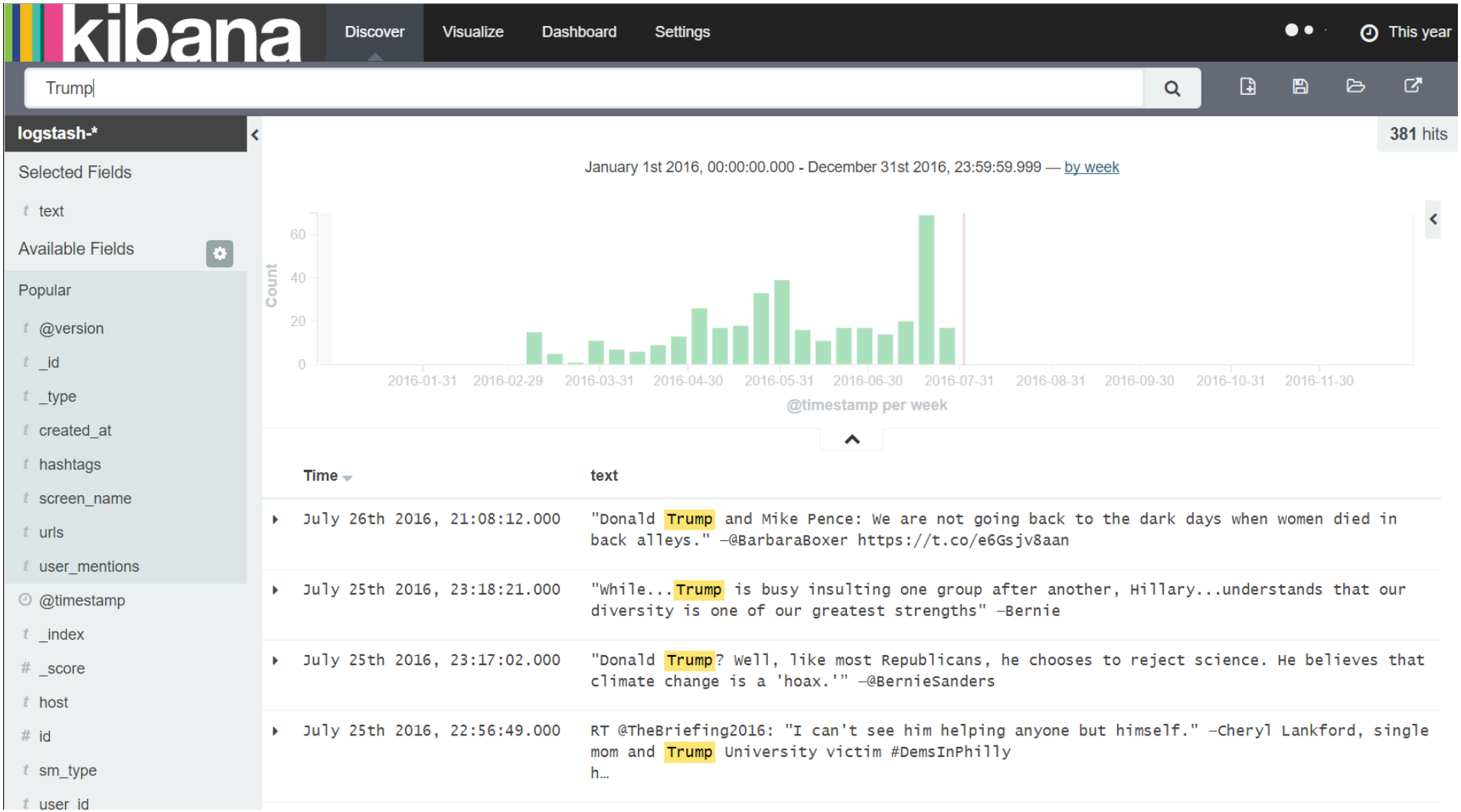

kibana
Discover
Visualize
Dashboard
Settings
This year
Search...
logstash-*
Selected Fields
text
Available Fields
Popular
@version
_id
_type
created_at
hashtags
screen_name
urls
user_mentions
@timestamp
_index
_score
host
id
sm_type
user_id
3,210 hits
January 1st 2016, 00:00:00.000 - December 31st 2016, 23:59:59.999 — by week
Count
250
200
150
100
50
0
2016-01-31
2016-02-29
2016-03-31
2016-04-30
2016-05-31
2016-06-30
2016-07-31
2016-08-31
2016-09-30
2016-10-31
2016-11-30
@timestamp per week
Time
text
July 27th 2016, 12:34:54.000 RT @TheBriefing2016: A few thoughts from @FLOTUS that seem particularly relevant right now. https://t.co/FgRTjWYUjl
July 27th 2016, 12:10:04.000 RT @timkaine: .@realDonaldTrump the Hotel Roanoke AC is fine & you should pay them for it. You're the one who brought the hot air! https://…
July 27th 2016, 11:47:52.000 History made. https://t.co/JRgDbFms3n
July 27th 2016, 10:52:50.000 "What I admire most about Hillary is that she never buckles under pressure." —@FLOTUS https://t.co/mbg5H8QZA3 #DemsInPHL
July 26th 2016, 23:21:28.000 We just put the biggest crack in that glass ceiling yet." —Hillary https://t.co/mYkaLIv861

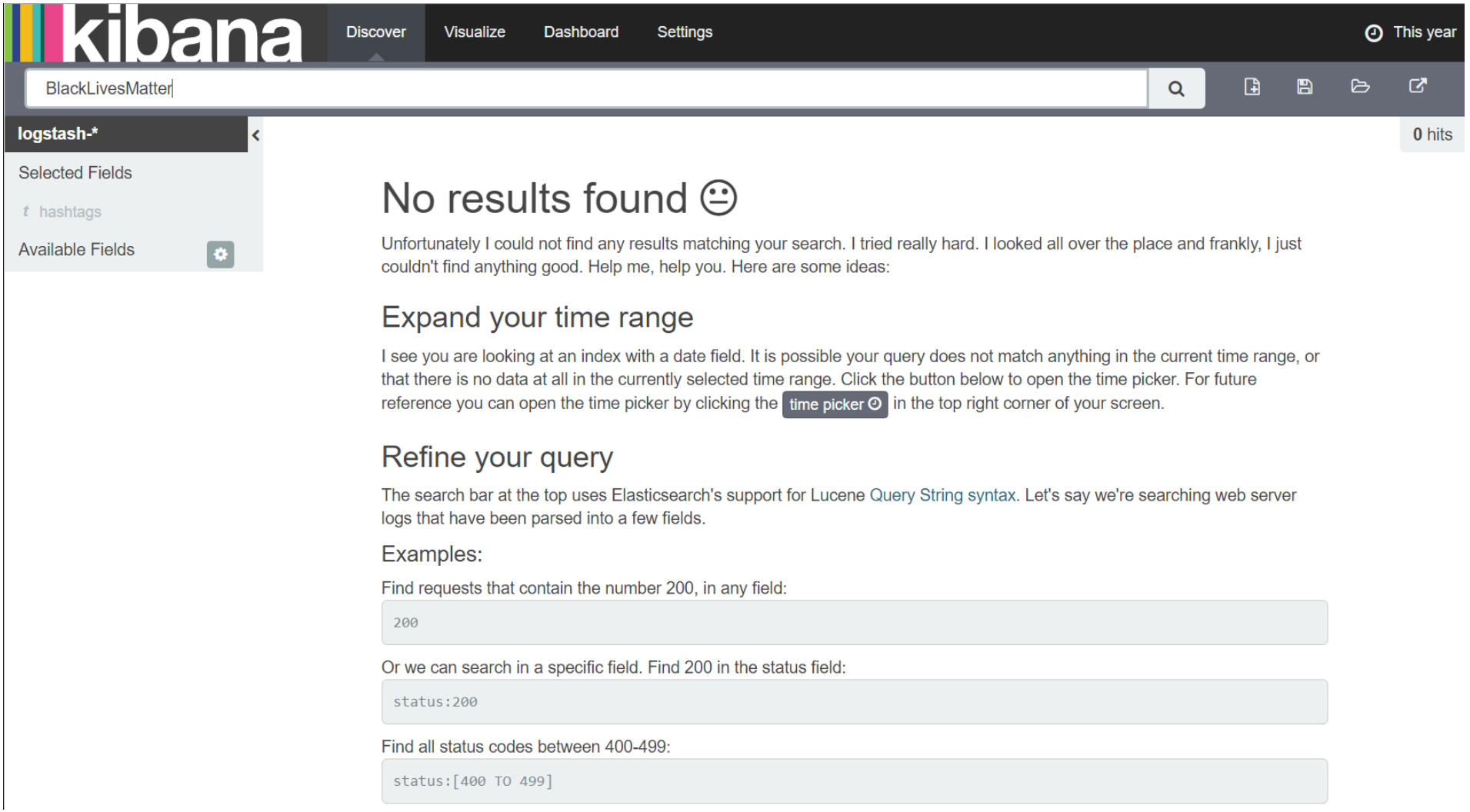



# Doc: logstash-2016.07.23/tweet/756860753668440000

Table | JSON

| | |
|---|---|
| @timestamp | July 23rd 2016, 10:37:20.000 |
| t @version | 1 |
| t _id | 756860753668440000 |
| t _index | logstash-2016.07.23 |
| # _score | 1 |
| t _type | tweet |
| t created_at | Sat Jul 23 14:37:20 +0000 2016 |
| t hashtags | |
| t host | c53a8a1f6b9b |
| # id | 756,860,753,668,440,064 |
| t screen_name | HillaryClinton |
| t sm_type | tweet |
| t text | This is what dauntless leadership looks like. Thank you, @GovernorVA—proud to call you a friend. -H https://t.co/c72c0HUSyw |
| t urls | https://twitter.com/NBC29/status/756641190511083523 |
| t user_id | 1339835893 |
| t user_mentions | GovernorVA |

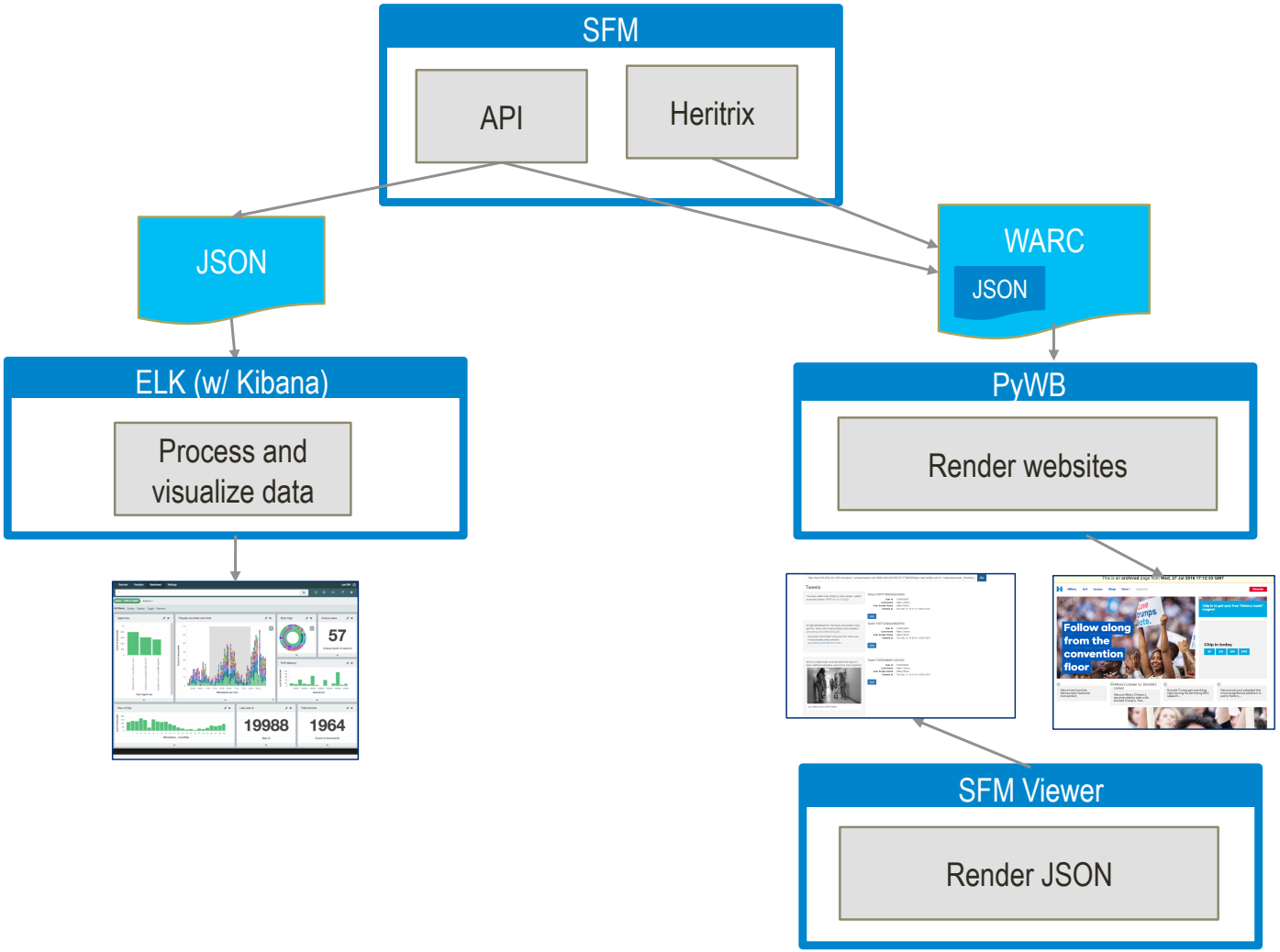

kibana
Discover
Visualize
Dashboard
Settings
This year
Trump
logstash-*
Selected Fields
text
Available Fields
Popular
@version
_id
_type
created_at
hashtags
screen_name
urls
user_mentions
@timestamp
_index
_score
host
id
sm_type
user_id
381 hits
January 1st 2016, 00:00:00.000 - December 31st 2016, 23:59:59.999 — by week
Count
60
40
20
0
2016-01-31
2016-02-29
2016-03-31
2016-04-30
2016-05-31
2016-06-30
2016-07-31
2016-08-31
2016-09-30
2016-10-31
2016-11-30
@timestamp per week
Time
text
July 26th 2016, 21:08:12.000
"Donald Trump and Mike Pence: We are not going back to the dark days when women died in back alleys." —@BarbaraBoxer https://t.co/e6Gsjv8aan
July 25th 2016, 23:18:21.000
"While...Trump is busy insulting one group after another, Hillary...understands that our diversity is one of our greatest strengths" —Bernie
July 25th 2016, 23:17:02.000
"Donald Trump? Well, like most Republicans, he chooses to reject science. He believes that climate change is a 'hoax.'" —@BernieSanders
July 25th 2016, 22:56:49.000
RT @TheBriefing2016: "I can't see him helping anyone but himself." —Cheryl Lankford, single mom and Trump University victim #DemsInPhilly
h…


kibana
Discover
Visualize
Dashboard
Settings
This year
BlackLivesMatter
logstash-*
Selected Fields
t hashtags
Available Fields
0 hits
No results found 😐
Unfortunately I could not find any results matching your search. I tried really hard. I looked all over the place and frankly, I just couldn't find anything good. Help me, help you. Here are some ideas:
Expand your time range
I see you are looking at an index with a date field. It is possible your query does not match anything in the current time range, or that there is no data at all in the currently selected time range. Click the button below to open the time picker. For future reference you can open the time picker by clicking the time picker in the top right corner of your screen.
Refine your query
The search bar at the top uses Elasticsearch's support for Lucene Query String syntax. Let's say we're searching web server logs that have been parsed into a few fields.
Examples:
Find requests that contain the number 200, in any field:
200
Or we can search in a specific field. Find 200 in the status field:
status:200
Find all status codes between 400-499:
status:[400 TO 499]

# ELK IS GREAT!
# (FOR USERS WHO DO DATA ANALYSIS)

## WHAT ABOUT OUR OTHER USERS?

- linked web content
- media (images, videos)
- browsing


SFM
API
Heritrix
JSON
WARC
JSON
ELK (w/ Kibana)
Process and visualize data
PyWB
Render websites
57
19988
1964
Follow along from the convention floor
SFM Viewer
Render JSON

# DEMO

SOCIAL FEED MANAGER (SFM)
ELK

PYWB
SFM VIEWER

(screenshots if web fails us)

# sfm Search Page

Search this collection by url:

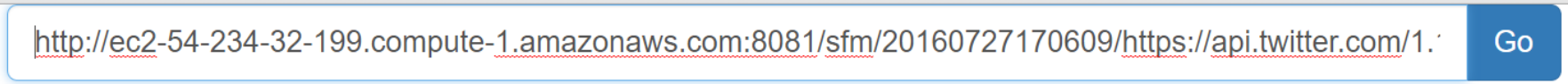
https://api.twitter.com/* Search

# pywb Query Results

**19** captures of **https://api.twitter.com/**

| Capture | Status | Original Url | Archive File |
|---|---|---|---|
| Wed, 27 Jul 2016 17:06:12 GMT | 200 | https://api.twitter.com/1.1/statuses/user_timeline.json?count=200&max_id=693862210079178752&user_id=1339835893 | f1e11c79d5aa42128ffc9005fa0aab39-20160727170606154-00000-32-acc6328176f6-8000.warc.gz |
| Wed, 27 Jul 2016 17:06:12 GMT | 200 | https://api.twitter.com/1.1/statuses/user_timeline.json?count=200&max_id=693954836040036351&user_id=1339835893 | f1e11c79d5aa42128ffc9005fa0aab39-20160727170606154-00000-32-acc6328176f6-8000.warc.gz |
| Wed, 27 Jul 2016 17:06:12 GMT | 200 | https://api.twitter.com/1.1/statuses/user_timeline.json?count=200&max_id=696075651086999551&user_id=1339835893 | f1e11c79d5aa42128ffc9005fa0aab39-20160727170606154-00000-32-acc6328176f6-8000.warc.gz |
| Wed, 27 Jul 2016 17:06:11 GMT | 200 | https://api.twitter.com/1.1/statuses/user_timeline.json?count=699317199027249151&user_id=1339835893 | f1e11c79d5aa42128ffc9005fa0aab39-20160727170606154-00000-32-acc6328176f6-8000.warc.gz |
| Wed, 27 Jul 2016 17:06:11 GMT | 200 | https://api.twitter.com/1.1/statuses/user_timeline.json?count=702961449560317951&user_id=1339835893 | f1e11c79d5aa42128ffc9005fa0aab39-20160727170606154-00000-32-acc6328176f6-8000.warc.gz |
| Wed, 27 Jul 2016 17:06:11 GMT | 200 | https://api.twitter.com/1.1/statuses/user_timeline.json?count=706654121114271743&user_id=1339835893 | f1e11c79d5aa42128ffc9005fa0aab39-20160727170606154-00000-32-acc6328176f6-8000.warc.gz |
| Wed, 27 Jul 2016 17:06:10 GMT | 200 | https://api.twitter.com/1.1/statuses/user_timeline.json?count=710286683296632831&user_id=1339835893 | f1e11c79d5aa42128ffc9005fa0aab39-20160727170606154-00000-32-acc6328176f6-8000.warc.gz |
| Wed, 27 Jul 2016 17:06:10 GMT | 200 | https://api.twitter.com/1.1/statuses/user_timeline.json?count=715703842411716608&user_id=1339835893 | f1e11c79d5aa42128ffc9005fa0aab39-20160727170606154-00000-32-acc6328176f6-8000.warc.gz |
| Wed, 27 Jul 2016 17:06:09 GMT | 200 | https://api.twitter.com/1.1/statuses/user_timeline.json?count=720795763316559872&user_id=1339835893 | f1e11c79d5aa42128ffc9005fa0aab39-20160727170606154-00000-32-acc6328176f6-8000.warc.gz |
| Wed, 27 Jul 2016 17:06:09 GMT | 200 | https://api.twitter.com/1.1/statuses/user_timeline.json?count=725151852262768639&user_id=1339835893 | f1e11c79d5aa42128ffc9005fa0aab39-20160727170606154-00000-32-acc6328176f6-8000.warc.gz |
| Wed, 27 Jul 2016 17:06:09 GMT | 200 | https://api.twitter.com/1.1/statuses/user_timeline.json?count=730800314417221632&user_id=1339835893 | f1e11c79d5aa42128ffc9005fa0aab39-20160727170606154-00000-32-acc6328176f6-8000.warc.gz |

[{"created_at":"Thu May 12 15:12:11 +0000 2016","id":730777600902242304,"id_str":"730777600902242304","text":"\"I\u2019ve been called many things by many people. 'Quitter' is not one of them.\" #TBT https:\/\/t.co\/wUjSOxc85I","truncated":false,"entities":{"hashtags":[{"text":"TBT","indices":[77,81]}],"symbols":[],"user_mentions":[],"urls":[{"url":"http://ec2-54-234-32-199.compute-1.amazonaws.com:8081/sfm/20160727170609/https:\/\/t.co\/wUjSOxc85I","expanded_url":"http://ec2-54-234-32-199.compute-1.amazonaws.com:8081/sfm/20160727170609/http:\/\/hrc.io\/1THGp03","display_url":"hrc.io\/1THGp03","indices":[82,105]}]},"source":"\u003ca href=\"http://ec2-54-234-32-199.compute-1.amazonaws.com:8081/sfm/20160727170609/https:\/\/about.twitter.com\/products\/tweetdeck\" rel=\"nofollow\"\u003eTweetDeck\u003c\/a\u003e","in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":1339835893,"id_str":"1339835893","name":"Hillary Clinton","screen_name":"HillaryClinton","location":"New York, NY","description":"Wife, mom, grandma, women+kids advocate, FLOTUS, Senator, SecState, hair icon, pantsuit aficionado, 2016 presidential candidate. Tweets from Hillary signed \u2013H","url":"http://ec2-54-234-32-199.compute-1.amazonaws.com:8081/sfm/20160727170609/https:\/\/t.co\/xhPHAcvdoc","entities":{"url":{"urls":[{"url":"http://ec2-54-234-32-199.compute-1.amazonaws.com:8081/sfm/20160727170609/https:\/\/t.co\/xhPHAcvdoc","expanded_url":"http:\/\/HillaryClinton.com","display_url":"HillaryClinton.com","indices":[0,23]}]},"description":{"urls":[]}},"protected":false,"followers_count":7773968,"friends_count":677,"listed_count":29060,"created_at":"Tue Apr 09 18:04:35 +0000 2013","favourites_count":1128,"utc_offset":-14400,"time_zone":"Eastern Time (US & Canada)","geo_enabled":true,"verified":true,"statuses_count":6954,"lang":"en","contributors_enabled":false,"is_translator":false,"is_translation_enabled":false,"profile_background_color":"0057B8","profile_background_image_url":"http://ec2-54-234-32-199.compute-1.amazonaws.com:8081/sfm/20160727170609/http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_image_url_https":"http://ec2-54-234-32-199.compute-1.amazonaws.com:8081/sfm/20160727170609/https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":false,"profile_image_url":"http://ec2-54-234-32-199.compute-1.amazonaws.com:8081/sfm/20160727170609/http:\/\/pbs.twimg.com\/profile_images\/750300510264107008\/G8-PA5KA_normal.jpg","profile_image_url_https":"http://ec2-54-234-32-199.compute-1.amazonaws.com:8081/sfm/20160727170609/https:\/\/pbs.twimg.com\/profile_images\/750300510264107008\/G8-PA5KA_normal.jpg","profile_banner_url":"http://ec2-54-234-32-199.compute-1.amazonaws.com:8081/sfm/20160727170609/https:\/\/pbs.twimg.com\/profile_banners\/1339835893\/1469281851","profile_link_color":"0057B8","profile_sidebar_border_color":"000000","profile_sidebar_fill_color":"000000","profile_text_color":"000000","profile_use_background_image":false,"has_extended_profile":true,"default_profile":false,"default_profile_image":false,"following":false,"follow_request_sent":false,"notifications":false},"geo":null,"coordinates":null,"place":null,"contributors":null,"is_quote_status":false,"retweet_count":809,"favorite_count":2117,"favorited":false,"retweeted":false,"possibly_sensitive":false,"lang":"en"},{"created_at":"Thu May 12 14:50:01 +0000 2016","id":730772020233928704,"id_str":"730772020233928704","text":"RT @TheBriefing2016: Paul Ryan and Donald Trump just met. Here's why it was probably pretty awkward.\nhttps:\/\/t.co\/Fxu0td8N78","truncated":false,"entities":{"hashtags":[],"symbols":[],"user_mentions":[{"screen_name":"TheBriefing2016","name":"The Briefing","id":3232054991,"id_str":"3232054991","indices":[3,19]}],"urls":[{"url":"http://ec2-54-234-32-199.compute-1.amazonaws.com:8081/sfm/20160727170609/https:\/\/t.co\/Fxu0td8N78","expanded_url":"http://ec2-54-234-32-199.compute-1.amazonaws.com:8081/sfm/20160727170609/https:\/\/amp.twimg.com\/v\/d901457a-a04c-443b-b363-f3643af4db9f","display_url":"amp.twimg.com\/v\/d901457a-a04\u2026","indices":[101,124]}]},"source":"\u003ca href=\"http://ec2-54-234-32-199.compute-1.amazonaws.com:8081/sfm/20160727170609/https:\/\/about.twitter.com\/products\/tweetdeck\" rel=\"nofollow\"\u003eTweetDeck\u003c\/a\u003e","in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":1339835893,"id_str":"1339835893","name":"Hillary Clinton","screen_name":"HillaryClinton","location":"New York, NY","description":"Wife, mom, grandma, women+kids advocate, FLOTUS, Senator, SecState, hair icon, pantsuit aficionado, 2016 presidential candidate. Tweets from Hillary signed \u2013H","url":"http://ec2-54-234-32-199.compute-1.amazonaws.com:8081/sfm/20160727170609/https:\/\/t.co\/xhPHAcvdoc","entities":{"url":{"urls":[{"url":"http://ec2-54-234-32-199.compute-1.amazonaws.com:8081/sfm/20160727170609/https:\/\/t.co\/xhPHAcvdoc","expanded_url":"http:\/\/HillaryClinton.com","display_url":"HillaryClinton.com","indices":[0,23]}]},"description":{"urls":[]}},"protected":false,"followers_count":7773968,"friends_count":677,"listed_count":29060,"created_at":"Tue Apr 09 18:04:35


http://ec2-54-234-32-199.compute-1.amazonaws.com:8081/sfm/20160727170609/https://api.twitter.com/1.
Go

http://ec2-54-234-32-199.compute-1.amazonaws.com:8081/sfm/20160727170609/https://api.twitter.com/1.1 Go

# Tweets

"I've been called many things by many people. 'Quitter' is not one of them." #TBT hrc.io/1THGp03

## Tweet 730777600902242300

| | |
|---|---|
| **User Id** | 1339835893 |
| **User Name** | Hillary Clinton |
| **User Screen Name** | HillaryClinton |
| **Created at** | Thu May 12 15:12:11 +0000 2016 |

Json

RT @TheBriefing2016: Paul Ryan and Donald Trump just met. Here's why it was probably pretty awkward. amp.twimg.com/v/d901457a-a04…

> Paul Ryan and Donald Trump just met. Here's why it was probably pretty awkward. amp.twimg.com/v/d901457a-a04…

## Tweet 730772020233928700

| | |
|---|---|
| **User Id** | 1339835893 |
| **User Name** | Hillary Clinton |
| **User Screen Name** | HillaryClinton |
| **Created at** | Thu May 12 14:50:01 +0000 2016 |

Json

80% of a child's brain is formed before the age of 3. Early childhood education could not

## Tweet 730762680911401000

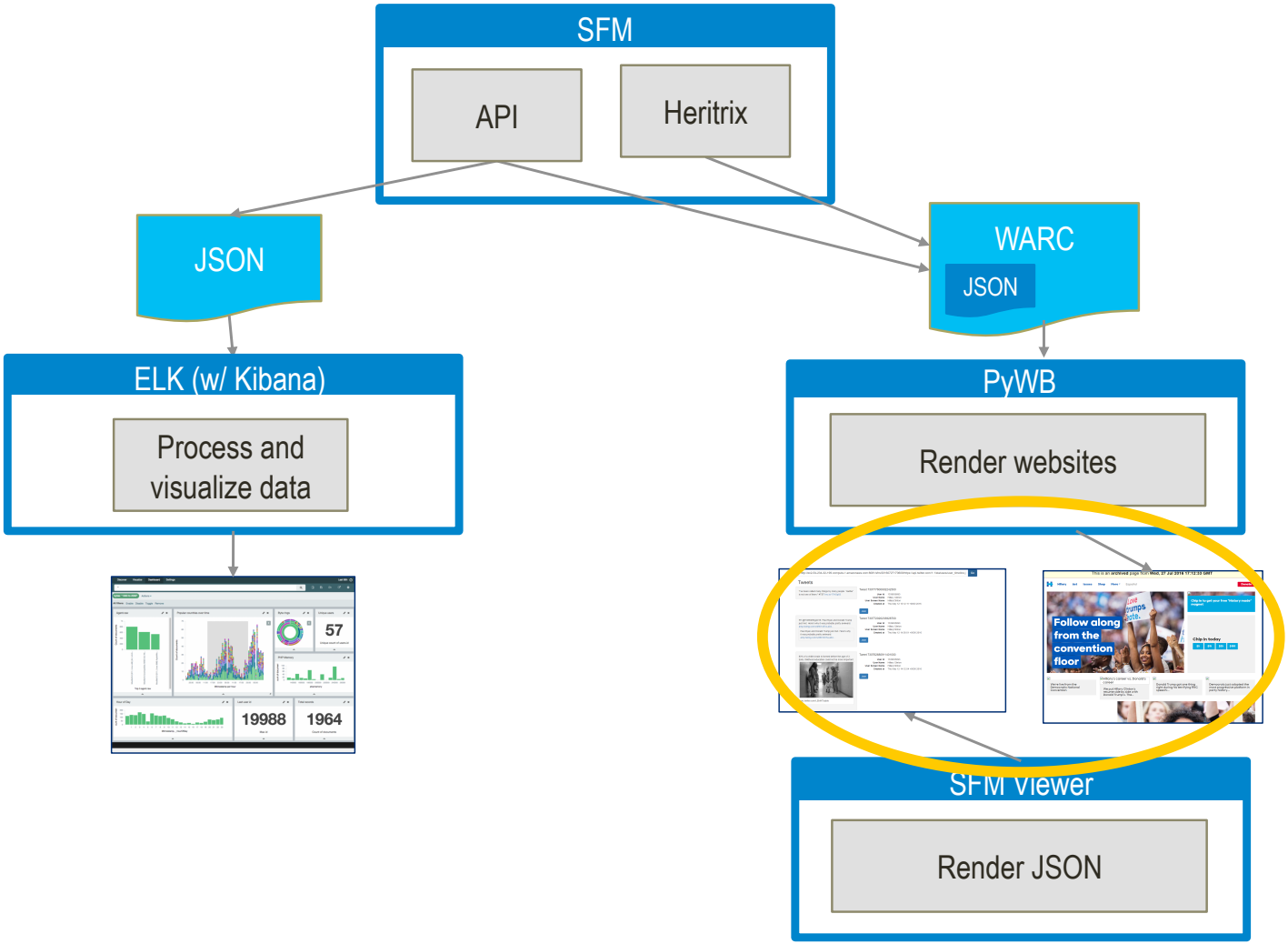
| | |
|---|---|
| **User Id** | 1339835893 |

# What these old photos and videos can tell you about a Hillary Clinton presidency

Look back at where she's been to see where she's going.

May 12, 2016 by John Buysse




SFM
API
Heritrix
JSON
WARC
JSON
ELK (w/ Kibana)
Process and visualize data
PyWB
Render websites
57
19988
1964
Follow along from the convention floor
SFM viewer
Render JSON

# WHAT'S THE TAKEAWAY?

technologically

It's possible to collect messy-web & social-media-web

. . . without creating access silos

. . . using existing technologies

. . . to make more users happy ☺ ☺ ☺