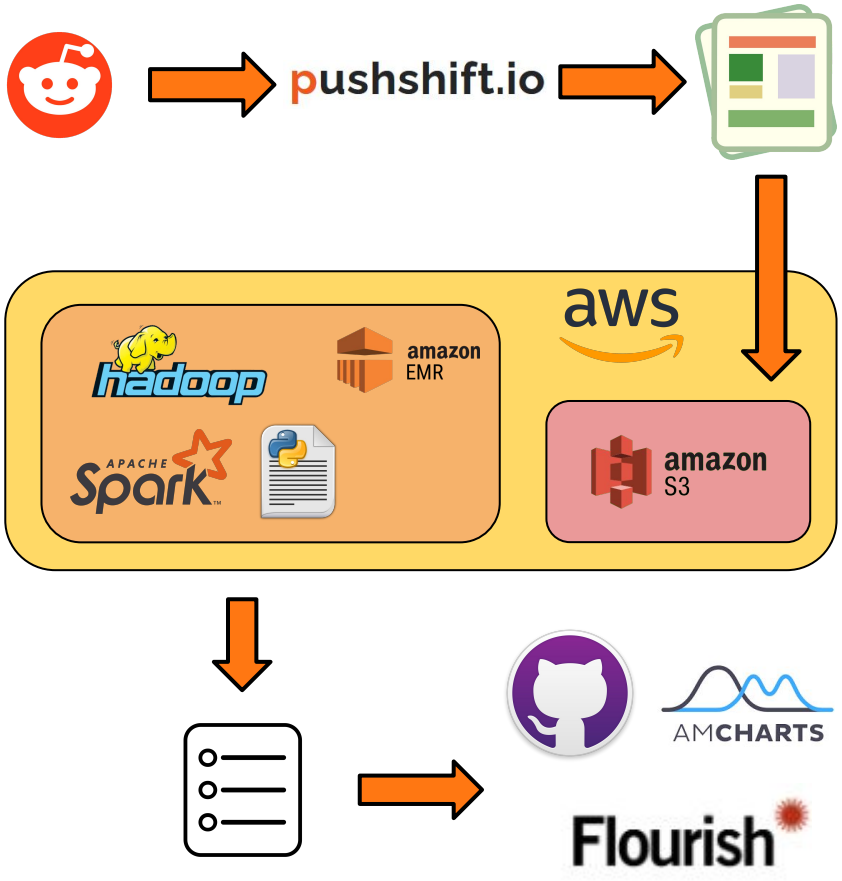
Análisis de reddit

Francisco Javier Lozano Hernández
Jorge Roselló Martín
Daniel Alcázar Muñoz

¿Que hemos hecho?



- Dataset extraído de Pushift.io
- Amazon Web Services: S3 y EMR
- Procesamiento de Big Data
- Ejecución del software
- Generamos el resultado
- Presentación en la web

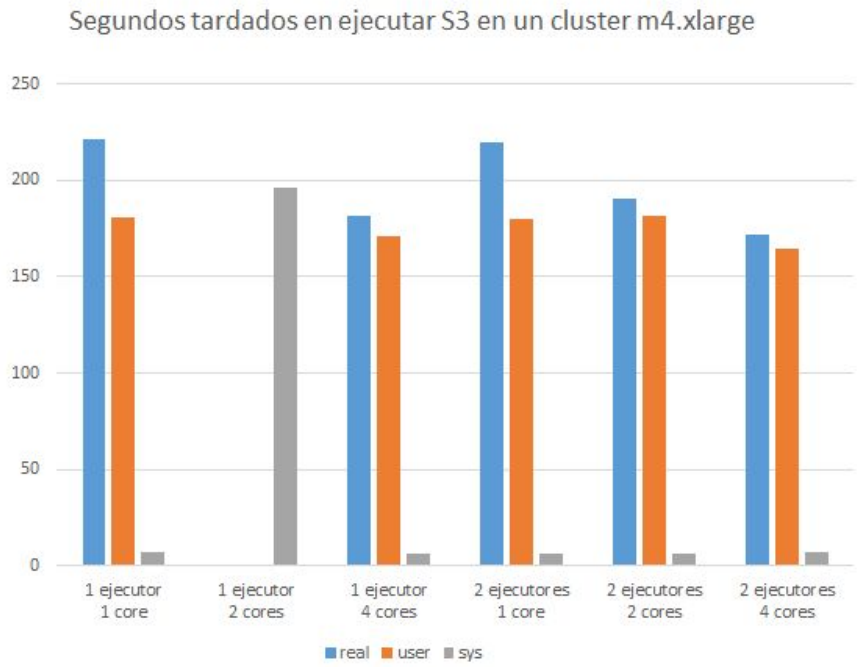
Breve resumen técnico

amazon
EMR


```
"name": "t3_abac2d",
"no_follow": false,
"num_comments": 71,
"num_crossposts": 5,
"num_reports": null,
"over_18": false,
"parent_whitelist_status": "all_ads",
"permalink": "/r/AnimalsBeingDerps/comments/abac2d/tiny_monkeys_eating_grapes/",
"pinned": false,
"post_hint": "link",
"preview": {
  "enabled": false,
  "images": [
    {
      "id": "kGc3H9QqRWRcVRmA4rLropzEE3g0LXSL3qaaVZPgsfc",
      "resolutions": [
        {
          "height": 108,
          "url": "https://external-preview.redd.it/RYWqrnfH5LwBa4t8RCcXtMZm-IFKCW4P
          "width": 108
        },
        {
```


- Clúster m4.xlarge
- Hadoop 2.8.5
- Spark 2.4.4
- Modelo de los datos
- Dataframes

Problemas



- 50 Gb en un instancia EC2, Spark crasheaba
- Lentitud general de ejecución
- Falta de memoria para manejar tantos datos

Visualización

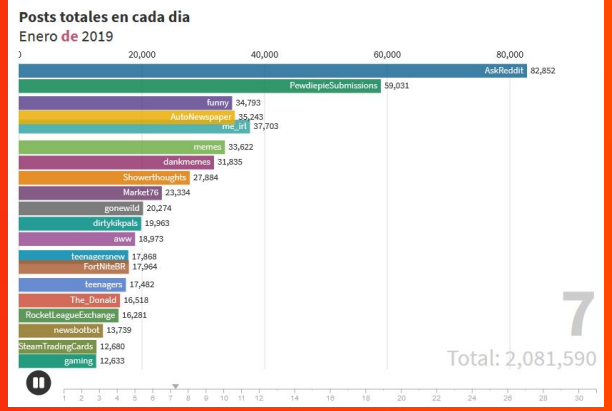
Página del proyecto realizada mediante GitHub Pages

AMCHARTS

Flourish

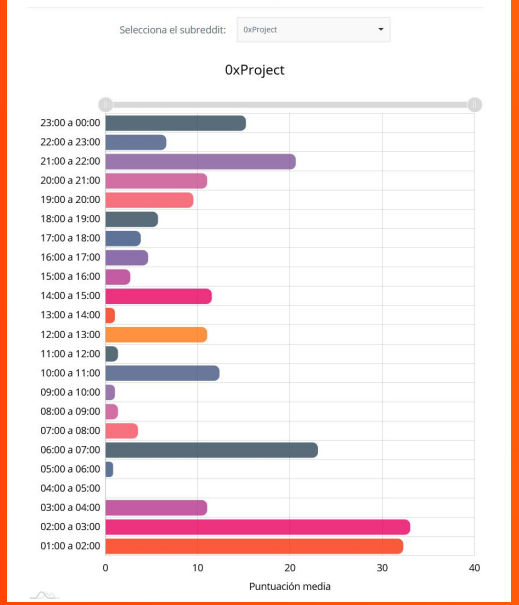
Script S3

Número de posts por dia y por subreddit



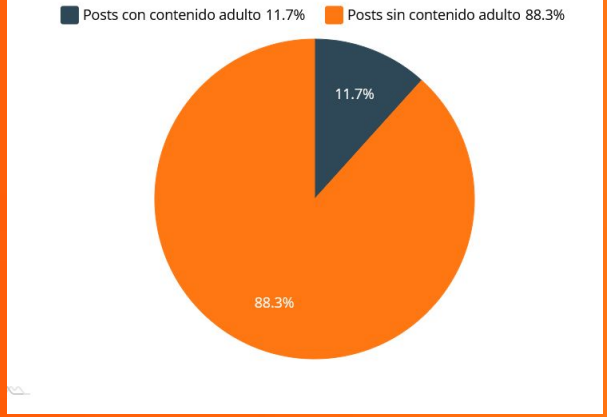
Script S2

Franja horaria (UTC) donde se consigue mayor puntuación en cada subreddit



Script S6

Número de posts en todo Reddit etiquetado cómo nsfw (Mayor de 18 años)



Si quieres ver más hecha un vistazo a nuestra página

https://beybo.github.io/ProyectoRedditCloud/

FIN

¿Preguntas?