

Reddit crawler

Περιβολαροπουλος Χρηστος
Πετρογιαννης Γιωργος

Τι είναι το Reddit?

“The front page of the internet”

-Reddit

Ισως –απλουστευμενα- ενα τεραστιο forum

- Επικαιροτητα
- Μουσικη
- Φωτογραφιες
- Προγραμματισμος
- ...

Πως τα βρισκουμε ολα αυτα

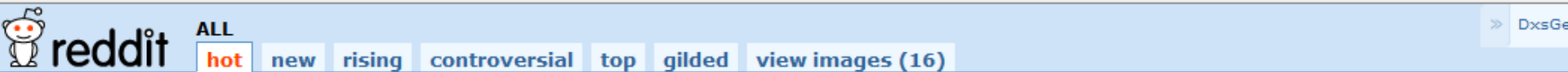
Sub-reddits: θεματικες ενοτητες

Subscribe

Frontpage

Your frontpage

MY SUBREDDITS - DASHBOARD - FRONT - **ALL** - RANDOM - FRIENDS | add shortcuts from the my subreddits menu at left or click the button by the subreddit name, drag and drop to sort



- 1



[Enter the D'Addario XL & Guitar Moves sweepstakes for a chance to win a Gretsch White Falcon guitar, a trip to NY, and more!](#) (promotions.daddario.com)
promoted by redditads
26 comments share save hide report

sponsored link what's this?
- 1 2322



[USA drops case against Wikileaks founder Julian Assange](#) Misleading title (smh.com.au)
(3587|1360) submitted 4 hours ago by kismor to worldnews
321 comments share save hide report [I+c]
- 2 1984



[Miley Cyrus - Wrecking Ball \(Chatroulette Version\) - \[2:45\]](#) (youtube.com)
 (3119|1459) submitted 4 hours ago by kismor to music
283 comments share save hide report [I+c]

YouTube title: Miley Cyrus - Wrecking Ball (Chatroulette Version)
- 3 3922



[UK Prime Minister David Cameron Announces That Filters Used to Block Porn Will Also Block Websites Espousing "Extremist" Views in Order "to Keep Our Country Safe"](#) (publications.parliament.uk)
(11849|7923) submitted 8 hours ago by MoonMetropolis to worldnews
2212 comments share save hide report [I+c]
- 4 2392




[God damn it Frank](#) (i.imgur.com)
 (4815|2500) submitted 5 hours ago by catherinsarris to gifs
114 comments share save hide report [I+c]
- 5 3374



[Water balloons half frozen with candles placed in the hollow center](#) (imgur.com)
 (16138|12809) submitted 9 hours ago by Proteon to pics
352 comments share save hide report [I+c]
- 6 2333



[As I looked through my in-laws depression-era family photos they told me to take what I wanted. This is the one I took.](#) (imgur.com)
 (3176|1100) submitted 8 hours ago by swejmar to pics
48 comments share save hide report [I+c]
- ↑



[After saving up for months I was finally able to buy my 9 year old arthritic Shepherd a huge posturepedic](#)

The “problem”

- πολλοι χρηστες - πολλα subreddits
- Ενδαιφερονταπραγματα για τα οποια δε γνωριζουμε
- Δυσκολο να τα ανακαλυψεις με την τωρινη οργανωση του reddit

The “idea”

Post το οποίο μ αρεσει

ο uploader μπορεί να εχει
κι αλλα ενδιαφεροντα ποστ

οι χρηστες που σχολιασαν
μπορει να εχουν
κι αλλα ενδιαφεροντα ποστ

Πως θα βρω αυτα τα post
γρηγορα?

The “idea”

- Αναδρομική εκτέλεση της παραπάνω αναζήτησης
- Δίκτυο (γραφός) απο post
- Καταταξη

How it's done

- Clojure api for reddit (reddit.clj)
- Ευκολη στη χρηση (50-60 γραμμες)
- Επιτρεπει χρηση αλλων εργαλειων (storm)

The solution's problem

- Too many posts
- Πολλή επεξεργαστική ισχύς
- Μετα απο μεγάλο αριθμο post αργή επεξεργασία

The solution's problem solution

Storm!

- Παραλληλη επεξεργασία για το twitter
- Json based
- Ευκολια στη χρηση
- Fault tolerant
- Guarantee of message processing

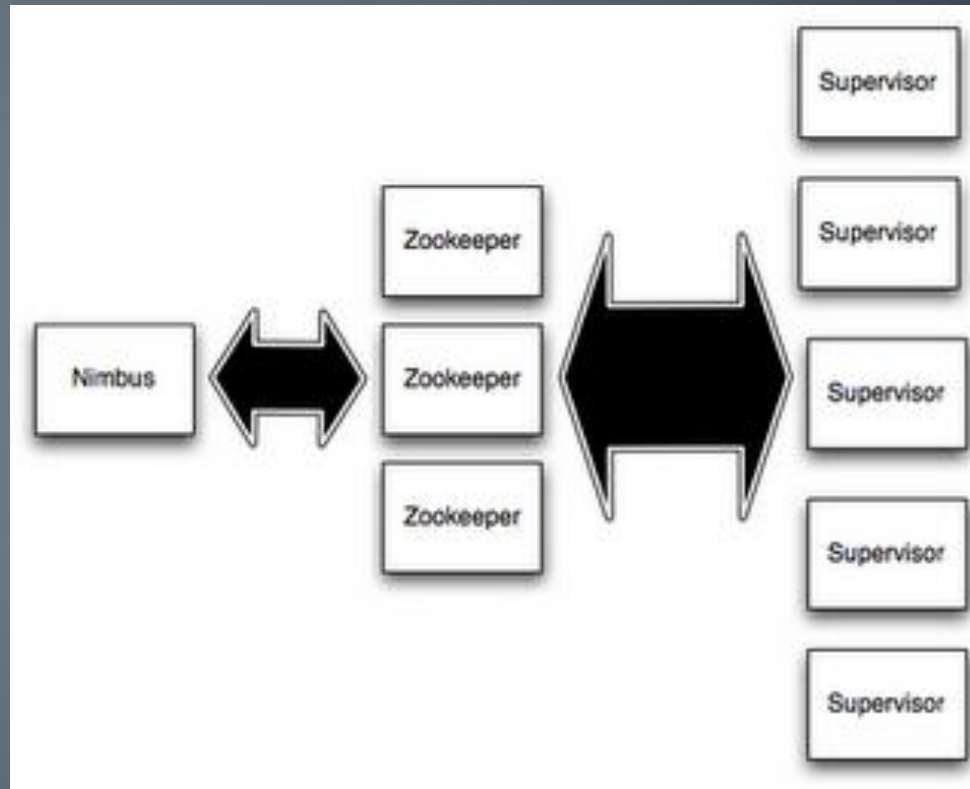
How it works

- Ομοιοτητα με το Hadoop
- Topologies
- Επεξεργάζεται post μεχρι να το “σκοτωσεις”

Storm hierarchy

- Master node: Nimbus
- Supervisor
- Connected by zookeepers

Storm hierarchy



Topologies

3 βασικοί οροι:

- Streams
- Spouts
- Bolts

Streams

- Ροη απο δεδομενα
- Just posts > interesting posts

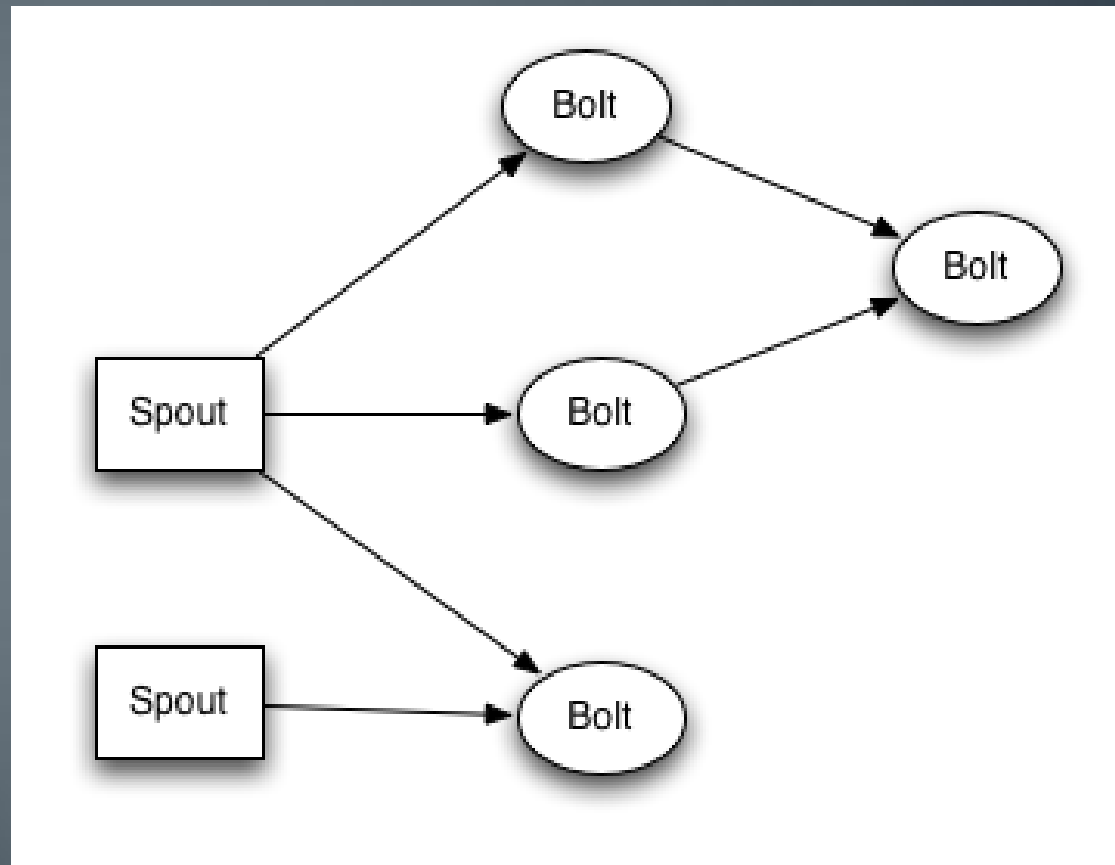
Spouts

- “Βρύση” των streams
- Αρχικός κομβος που παρέχει δεδομένα – posts

Bolts

- Τροφοδοτείται απο τα sprouts
- Λαμβανει streams και τα επεξεργάζεται
- Αποστολη streams σε αλλα bolts

Topology



Topology

Καθε `spout` τροφοδοτει με ενα `stream` απο `posts` τα `bolts` με τα οποια ειναι συνδεδεμενο

Τα `bolts` επεξεργαζονται τα `streams` και επειτα στελνουν σε αλλα `bolts` για αλλη επεξεργασία

Καθε κομβος μπορεί να εκτελεσει επεξεργασία παραλληλα με τους υπολοιπους

Links

- <https://github.com/nathanmarz/storm>
- <https://github.com/fakedrake/reddit-crawler>