Προγραμματισμός & Συστήματα στον Παγκόσμιο Ιστό
Εργαστηριακή Άσκηση Ακαδημαϊκού Έτους 2020 - 2021

Θέμα: Σύστημα πληθοποριστικής συλλογής και ανάλυσης δεδομένων κίνησης HTTP

Στόχος

Στόχος της παρούσας εργασίας είναι η ανάπτυξη ενός **πλήρους συστήματος συλλογής, διαχείρισης και ανάλυσης πληθοποριστικής (crowdsourced) πληροφορίας,** που αφορά δεδομένα κίνησης HTTP.

Λειτουργικές Προδιαγραφές

Η κίνηση στο διαδίκτυο μέσω HTTP, μπορεί να καταγραφεί από οποιονδήποτε πελάτη (client), ώστε τα δεδομένα αυτά να χρησιμοποιηθούν αργότερα για την ανάλυση της συμπεριφοράς ενός ιστοτόπου. Για το σκοπό αυτό, έχει δημιουργηθεί το πρότυπο HAR (H*TTP* AR*chive*), που ορίζει μια συγκεκριμένη δομή (σχήμα) JSON για την αποθήκευση αυτών των δεδομένων. Λεπτομέρειες για το σχήμα μπορείτε να δείτε εδώ http://www.softwareishard.com/blog/har-12-spec/

Η καταγραφή δεδομένων HAR μπορεί να γίνει από τα developer tools δημοφιλών browsers ή 3rd party εργαλείων, π.χ.

- Chrome:

  https://developers.google.com/web/tools/chrome-devtools/network/reference#top_of_page
- Firefox:

  https://developer.mozilla.org/en-US/docs/Tools/Network_Monitor/Toolbar
- HTTP Toolkit: https://httptoolkit.tech/

Αν και η ανάλυση των δεδομένων ενός υπολογιστή δεν έχει τόσο μεγάλο ενδιαφέρον, η ανάλυση HAR αρχείων από πολλούς υπολογιστές, που αφορούν πολλούς ιστότοπους και διαφορετικές ώρες πρόσβασης, έχει τη δυνατότητα να αποκαλύψει ιδιαίτερα ενδιαφέρουσες πτυχές της υποδομής του παγκόσμιου ιστού. Έτσι λοιπόν, σκοπός της εργασίας είναι να κατασκευαστεί ένα σύστημα πληθοποριστικής συλλογής δεδομένων HAR με σκοπό την παροχή κάποιων βασικών αναλύσεων για κάθε χρήστη ξεχωριστά, αλλά και γενικότερων αναλύσεων που αφορούν την υποδομή διαδικτύου σε μια περιοχή (π.χ. Πάτρα). Στο σύστημα υπάρχουν δύο τύποι χρηστών: Διαχειριστής και Χρήστης.

Χρήστης

Ο χρήστης συνδέεται στο σύστημα μέσω σταθερού υπολογιστή, και έχει τις εξής δυνατότητες:

1) **Εγγραφή στο σύστημα**. Ο χρηστής εγγράφεται και αποκτά πρόσβαση στο σύστημα επιλέγοντας κάποιο username & password της αρεσκείας του, και παρέχοντας το email του. Το password πρέπει να είναι τουλάχιστον 8 χαρακτήρες και να περιέχει τουλάχιστον ένα κεφαλαίο γράμμα, ένα αριθμό και κάποιο σύμβολο (π.χ. #$*&@).
2) **Upload δεδομένων**. Ο χρήστης επιλέγει ένα αρχείο HAR από τον υπολογιστή του. Το αρχείο θα επεξεργαστεί τοπικά για την απαλοιφή ευαίσθητων δεδομένων και στη συνέχεια ο χρήστης έχει δύο επιλογές: α) Να το ανεβάσει στο σύστημα ή β) Να αποθηκεύσει το επεξεργασμένο αρχείο τοπικά.

Στην περίπτωση ανεβάσματος του αρχείου στο σύστημα, θα πρέπει να γίνει περαιτέρω επεξεργασία (στο server) των δεδομένων που θα ανέβουν, ώστε να αποθηκευτούν τα επιθυμητά στοιχεία με την κατάλληλη μορφή. Επίσης, θα πρέπει να «αναλυθεί» η IP του χρήστη που ανεβάζει το αρχείο, ώστε να ανακαλύπτεται αυτόματα ο πάροχος συνδεσιμότητας του χρήστη και να αποθηκεύεται η πληροφορία αυτή στη βάση μαζί με τις εγγραφές.

3) **Διαχείριση προφίλ**. Ο χρήστης μπορεί να αλλάξει το username/password και να δει βασικά στατιστικά για τα δεδομένα που έχει ανεβάσει (ημερομηνία τελευταίου upload, πλήθος εγγραφών)

4) **Οπτικοποίηση δεδομένων**. Ο χρήστης μπορεί να δει σε χάρτη τις τοποθεσίες των IPs στις οποίες έχει αποστείλει HTTP αιτήσεις. Συγκεκριμένα, στο χάρτη εμφανίζεται ένα heatmap που απεικονίζει την κατανομή του πλήθους των εγγραφών που αφορούν ιστοαντικείμενα τύπου HTML, PHP, ASP, JSP (ή σκέτα domains, χωρίς path).

Διαχειριστής

Ο Διαχειριστής αποκτά πρόσβαση στο σύστημα με σταθερό υπολογιστή, μέσω κατάλληλου μηχανισμού username / password. Κατά την είσοδό του στο σύστημα έχει τις εξής δυνατότητες.

1. **Απεικόνιση Βασικών Πληροφοριών.** Ο διαχειριστής βλέπει σε μία σελίδα κατάλληλη πληροφορία, σε πίνακες ή/και γραφήματα τα οποία απεικονίζουν:
    a. Το πλήθος των εγγεγραμμένων χρηστών
    b. Το πλήθος των εγγραφών στη βάση ανά τύπο (μέθοδο) αίτησης
    c. Το πλήθος των εγγραφών στη βάση ανά κωδικό (status) απόκρισης
    d. Το πλήθος των μοναδικών domains που υπάρχουν στη βάση
    e. Το πλήθος των μοναδικών παρόχων συνδεσιμότητας που υπάρχουν στη βάση
    f. Τη μέση ηλικία των ιστοαντικειμένων τη στιγμή που ανακτήθηκαν, ανά CONTENT-TYPE
2. **Ανάλυση χρόνων απόκρισης σε αιτήσεις (αντικείμενο τύπου `entries`, πεδίο `timings`)**. Εμφανίζεται παραμετροποιήσιμο διάγραμμα με το μέσο χρόνο απόκρισης (άξονας Υ) σε κάθε αίτηση ανά ώρα της ημέρας [0-24] (άξονας Χ). Το διάγραμμα μπορεί να απεικονίζει φιλτραρισμένα δεδομένα ως εξής:
    a. Είδος ιστοαντικειμένου (επιλογή ενός ή περισσότερων CONTENT-TYPE ή όλα)
    b. Ημέρα της εβδομάδας (Δευτέρα – Κυριακή ή όλα)
    c. Είδος HTTP μεθόδου κατά την αίτηση (επιλογή μιας ή περισσότερων, ή όλες)
    d. Πάροχος συνδεσιμότητας (π.χ. “Wind”, “Cosmote” ή όλα)
3. **Ανάλυση κεφαλίδων HTTP (αντικείμενα τύπου `headers`)**. Ο διαχειριστής βλέπει σε μία σελίδα κατάλληλη πληροφορία, σε πίνακες ή/και γραφήματα τα οποία απεικονίζουν στοιχεία που αφορούν τη χρήση κρυφών μνημών. Πιο συγκεκριμένα:
    a. Ιστόγραμμα κατανομής των TTL των ιστοαντικειμένων στην απόκριση, ανά CONTENT-TYPE (επιλογή ενός ή περισσότερων CONTENT-TYPE ή όλα). Το TTL είναι το max-age directive (αν υπάρχει) ή υπολογίζεται με βάση το expires (αν υπάρχει) και την ημερομηνία τροποποίησης του ιστοαντικειμένου. Το πλήθος των buckets του ιστογράμματος είναι 10 και το εύρος του κάθε bucket υπολογίζεται δυναμικά ανάλογα με τις ανακτώμενες τιμές.
    b. Ποσοστό max-stale και min-fresh directives επί του συνόλου των αιτήσεων ανά CONTENT-TYPE (επιλογή ενός ή περισσότερων CONTENT-TYPE ή όλα).
    c. Ποσοστό cacheability directives (public, private, no-cache, no-store) επί του συνόλου των αποκρίσεων ανά CONTENT-TYPE (επιλογή ενός ή περισσότερων CONTENT-TYPE ή όλα).

Όλα τα ανωτέρω γραφήματα/πίνακες παραμετροποιούνται με την επιλογή του πάροχου συνδεσιμότητας Πάροχος συνδεσιμότητας (π.χ. “Wind”, “Cosmote” ή όλα)

4. **Οπτικοποίηση δεδομένων (αντικείμενο `entries`)**. Ο διαχειριστής μπορεί να δει σε χάρτη τις τοποθεσίες των IPs στις οποίες έχει αποστείλει HTTP αιτήσεις. Συγκεκριμένα, εμφανίζεται ένας δείκτης (marker) ανά IP, με γραμμές που συνδέουν την πόλη προέλευσης του κάθε χρήστη με κάθε εικονίδιο. Το πάχος των γραμμών προσαρμόζεται ανάλογα με το πλήθος των αιτήσεων που έχουν γίνει προς τη συγκεκριμένη IP, κανονικοποιημένο ως προς το μέγιστο πλήθος που έχουν γίνει στη δημοφιλέστερη IP.

Περιορισμοί

1. Ομάδες 3 (τριών) το πολύ ατόμων.
2. Ελεγχόμενη πρόσβαση στα υποσυστήματα που απαιτούν σύνδεση/ αποσύνδεση.
3. Οι τεχνολογίες που θα χρησιμοποιηθούν θα είναι από τις διδαχθείσες στο μάθημα. Μπορείτε όμως να χρησιμοποιήσετε **επιπλέον** όποια άλλη τεχνολογία κρίνετε απαραίτητο αρκεί να είναι open-source.
4. Η εμφάνιση και η λειτουργικότητα της εφαρμογής αξιολογείται.

Παραδοτέα

1. Συνοπτική αναφορά που θα περιλαμβάνει τον σχεδιασμό της βάσης (ER, σχέσεις πινάκων).
2. Τον πηγαίο κώδικα και ένα export της ΒΔ

Παράρτημα

Προστασία ιδιωτικότητας

Τα δεδομένα HAR που επιλέγει ο χρήστης θα πρέπει να επεξεργάζονται τοπικά στον client (browser) με χρήση Javascript, πρωτού «ανέβουν» προς το server που περιέχει τη βάση δεδομένων. Αυτό είναι ιδιαίτερα σημαντικό καθώς τα δεδομένα HAR περιέχουν πολύ ευαίσθητες πληροφορίες του χρήστη (π.χ. form data, cookies κλπ). Θα πρέπει να διασφαλίσετε πως τα δεδομένα που αποστέλλονται είναι εντελώς ανώνυμα και δεν περιέχουν καμία πρόσθετη πληροφορία. Για την ανάγνωση των αρχείων σε JavaScript θα πρέπει να χρησιμοποιήσετε το FILE API της HTML5 (π.χ. https://www.html5rocks.com/en/tutorials/file/dndfiles/)

ΠΡΟΣΟΧΗ: project που δεν πληρούν τις κατάλληλες προδιαγραφές διασφάλισης των δεδομένων θα έχουν ιδιαίτερα δυσμενή βαθμολογική μεταχείριση.

Δεδομένα που τηρούνται στο server σας

Σύμφωνα με τα παραπάνω (αλλά και σύμφωνα με την εθνική και ευρωπαϊκή νομοθεσία), ο server σας θα πρέπει να συλλέγει ΜΟΝΟ τα απολύτως απαραίτητα δεδομένα για να μπορεί να επιτελέσει τις λειτουργίες χρήστη και διαχειριστή. Συνεπώς, με βάση το σχήμα HAR, κατά την τοπική επεξεργασία θα πρέπει να διατηρείται μόνο η πληροφορία που προέρχεται από τα παρακάτω πεδία:

Αντικείμενο	Πεδίο / Αντικείμενο
entries	startedDateTime
	timings
	serverIPAddress
timings	wait
request	method
	url (κρατάτε μόνο το domain)
	headers
response	status
	statusText
	headers
headers	content-type
	cache-control
	pragma
	expires
	age
	last-modified
	host

Επιπρόσθετα, κάποια από αυτή την πληροφορία θα πρέπει να χρησιμοποιήσετε ως βάση για να μπορέσετε να εξάγετε πρόσθετες πληροφορίες, όπως:

- Γεωγραφικές συντεταγμένες της IP του server που αποκρίνεται.
- Πάροχος συνδεσιμότητας του χρήστη κατά τη στιγμή του upload.
- Πόλη (συντεταγμένες) του χρήστη κατά τη στιγμή του upload.

Ανάκτηση χωρικής πληροφορίας από IP

Για τον εντοπισμό των γεωγραφικών συντεταγμένων στις οποίες αντιστοιχεί μια διεύθυνση IP θα πρέπει να χρησιμοποιηθεί κατάλληλο REST API. Ενδεικτικά, μπορείτε να μελετήσετε τα παρακάτω, τα περισσότερα από τα οποία λειτουργούν δωρεάν, με κάποιους περιορισμούς: ip-api, ipstack, ipapi, freegoip.

Συλλογή δεδομένων HAR

Τα μέλη της κάθε ομάδας θα πρέπει να συλλέξουν HAR αρχεία από τους υπολογιστές τους για πάραδειγμα, καταγράφοντας την κίνηση σε διάφορα browsing sessions επί ένα ικανό χρονικό διάστημα, τουλάχιστον μίας εβδομάδας. Αν κάποια ομάδα επιθυμεί, μπορεί να αξιοποιήσει τις automation δυνατότητες που έχουν εργαλεία όπως το HTTP Toolkit και να γράψει κάποιο script που διενεργεί συστηματική συλλογή δεδομένων

(π.χ. με χρήση curl σε προκαθορισμένη λίστα ιστοτόπων), ή χρησιμοποιώντας κάποια web crawling extensions.

Κάθε ομάδα θα πρέπει να δημιουργήσει στο σύστημά της τουλάχιστον 5 χρήστες και για κάθε ένα από αυτούς να περάσει στο σύστημα από ένα αρχείο δεδομένων. Κατά περίπτωση, επειδή το αρχείο δεδομένων που θα επιλέξετε για κάποιο χρήστη μπορεί να είναι αρκετά αραιό, μπορείτε να προσθέσετε κι άλλα αρχεία δεδομένων στον ίδιο χρήστη.

Χρήση τεχνολογιών

- Θα χρησιμοποιήσετε αποκλειστικά open-source τεχνολογίες για τη ΒΔ (π.χ. MySQL, PostgreSQL), τους χάρτες (Leaflet), τα γραφήματα (προτείνεται η chart.js) και τα heatmaps (https://leafletjs.com/plugins.html#heatmaps, προτείνεται η leaflet-heatmap.js).
- Η φόρτωση δεδομένων από τη ΒΔ θα πρέπει να γίνει χρησιμοποιώντας αποκλειστικά τεχνικές AJAX (προσοχή: όχι PHP για τη δημιουργία Javascript κώδικα).
- Τα δεδομένα HAR μπορεί να είναι αρκετά μεγάλα, ειδικά αν έχετε πολλά δεδομένα χρήσης. Για την επεξεργασία και καταχώρησή τους στη ΒΔ καλό είναι να σκεφτείτε τη χρήση ενός streaming parser (π.χ. salsify, json-machine).
- Προσοχή στη βάση δεδομένων: Χρησιμοποιήστε κατάλληλα indexes στους πίνακες ώστε να επιταχύνονται τα queries. Επίσης φροντίστε ώστε η είσοδος δεδομένων στη βάση να γίνεται με bulk inserts (π.χ. 1000 εγγραφές μαζί) και όχι μια-μια εγγραφή, για να επιταχύνεται η διαδικασία.