Zasebnost na prvem mestu - Arhitektura projekta BATMAN

Privacy by design - Architecture behind BATMAN project

Jani Bizjak
Institut Jožef Stefan,
Mednarodna podiplomska šola
Jožefa Stefana
Jamova cesta 39
1000 Ljubljana
jani.bizjak@ijs.si

Laura Guzelj Blatnik
Fakulteta za Matematiko in
Fiziko
Jadranska ulica 19
1000 Ljubljana
laura.g.blatnik@gmail.com

Nejc Mlakar
Fakulteta za Računalništvo in
Informatiko
Večna pot 113
1000 Ljubljana
nejcmlakar37@gmail.com

Anton Gradišek
Institut Jožef Stefan
Jamova cesta 39
1000 Ljubljana
anton.gradisek@ijs.si

Matjaž Gams
Institut Jožef Stefan,
Mednarodna podiplomska šola
Jožefa Stefana
Jamova cesta 39
1000 Ljubljana
matjaz.gams@ijs.si

POVZETEK

Zasebnost in varovanje osebnih podatkov ter njihovo izkoriščanje za namene oglaševanja postaja vse večji problem današnje družbe. Iz podatkov, ki se jih lahko pridobi iz različnih pametnih naprav - najbolj prednjači mobilni telefon, ki neprestano spremlja uporabnike - se lahko ustvari zanesljiv model, ki uporabnika v veliko primerih celo bolje "pozna", kot se pozna sam. To je še posebej kočljivo, ko govorimo o občutljivih medicinskih podatkih. V tem prispevku je opisana (arhitekturna) zasnova, ki se bo uporabljala pri projektu ERA PerMed BATMAN (ang. Biomolecular Analyses for Tailored Medicine in Acne iNversa), kjer bomo z integracijo različnih vrst podatkov odkrivali nova znanja o dermatološki bolezni.

Ključne besede

BATMAN, zasebnost, ERA PerMed, GDPR

ABSTRACT

Privacy and personal data exploitation are becoming a major concern in modern (Western) society. From the data obtained from different connected (smart) devices – smartphones being the biggest problems, as they tend to follow the user 24/7, machine learning and AI algorithms can model a person's behaviour in a way that it knows them better than the users knows themselves. Especially problematic are clinical data (health records). In this paper we describe an architectural design that is going to be used in the ERA PerMED project BATMAN (Biomolecular Analyses for Tailored Medicine in Acne iNversa) where we will integrate various types of data to discover new knowledge about the dermatological condition.

Keywords

BATMAN, privacy, ERA PerMed, GDPR

1. UVOD

Z večanjem priljubljenosti socialnih mrež, spleta ter mobilnega oglaševanja se hkrati veča tudi zanimanje podjetij za osebne podatke uporabnikov teh storitev. Trenutno se v večini primerov pridobljeni podatki uporabljajo za prilagojeno oglaševanje, zaradi količine in specifičnosti podatkov pa je le vprašanje časa, kdaj se bodo takšni podatki začeli zlorabljati na druge načine. Verjetno najbolj znan primer zadnjih let je t.i. *Facebook-Cambridge Analytica data scandal [4]*, kjer so se podatki o uporabnikih zlorabili za plenilsko oglaševanje, s katerimi naj bi naročniki vplivali na izide predsedniških volitev v ZDA.

Še večji problem nastane, ko so podatki, ki jih uporabniki delijo oz. se ti shranjujejo, medicinski. Ni si težko predstavljati hipotetičnega primera, kjer zavarovalnica, oborožena s tovrstnimi podatki, enostavno preračuna verjetnost za določen tip bolezni, ki jo ima zavarovanec pri njih, ter na ta način prilagodi (poviša) zavarovalno premijo zavarovanca, oz. ga v primerih previsokega tveganja sploh ne želi zavarovati.

Do neke mere to problematiko v Evropi rešuje zakon GDPR (General Data Protection Regulation, Splošna uredba EU o varstvu podatkov) [7], ki uporabniku omogoča vpogled v podatke, ki jih določeno podjetje hrani o njih, ter dovoljuje uporabnikom možnost zahteve za izbris. Zaradi vse pogostejših škandalov, kjer nepooblaščene osebe pridobijo dostop do osebnih podatkov uporabnikov in posledično zlorabe teh podatkov na takšen ali drugačen način, se zaupanje uporabnikov do sistemov, ki hranijo njihove podatke, zmanjšuje.

Na koncu to privede do tega, da uporabniki aplikacij ne želijo več uporabljati.

Enostavno je predvideti, da bo v prihodnje vedno več sistemov digitaliziranih, saj so tako učinkovitejši, bolj pregledni ter enostavnejši za uporabo (e-medicina, e-volitve, e-uprava...). Vendar je potrebno te sisteme že zdaj zasnovati tako, da bodo uporabniški podatki zaščiteni pred zlorabami, enostavni za vzdrževanje ter pregledni - le na takšen način jim bodo uporabniki zaupali in jih uporabljali.

V tem prispevku opisujemo primer arhitekturne rešitve za projekt ERA PerMed BATMAN. Projekt BATMAN se ukvarja z raziskovanjem bolezni *Acne Inversa* oz. *Hidradenitis suppurativa* [8] ter njenega zdravljenja. V sklopu projekta se bodo zbirali številni podatki o pacientih, od spremljanja aktivnosti z mobilno aplikacijo, izpolnjevanja anket, slik, pa do sekvenciranja DNK vzorcev pacientov. Pomembno je, da so podatki pravilno in varno shranjeni ter zaščiteni, tako da se tudi v primeru nezakonitega vdora v strežniški sistem ohranja anonimnost pacientov, hkrati pa je potrebno vpogled v podatke omogočiti raziskovalcem ter zdravnikom, ki bodo na podlagi zbranih podatkov dobili boljši vpogled v bolezen ter njeno zdravljenje.

2. ZASNOVA ZA ZASEBNOST

Pri javno dostopnih sistemih, še posebej takih, ki shranjujejo občutljive podatke, je pomembno, da se zasnova za varnost gradi na vseh korakih, tako pri strojni opremi, na kateri tečejo programi, kot v samih programih.

2.1 Strežniška arhitektura

Pogosto se dogaja, da so javno dostopne spletne strani deležne dnevnih napadov, v veliki večini kar avtomatiziranih agentov (t.i. botov) [2], ki iščejo morebitne ranljivosti v sistemih. Dobra praksa je ločitev strojne opreme na več podsistemov, kar poleg povečane varnosti nudi tudi dodatno varnost za shranjevanje podatkov, lažjo skalabilnost ter optimizacijo.

Strojna arhitektura v projektu BATMAN je ločena v tri samostojne sisteme: *1.* strežnik na kateri teče spletna stran, *2* strežnik namenjen izključno shranjevanju podatkov ter *3* nepovezan (ang. offline) strežnik za varnostne kopije (ang. backup), kjer se v določenih intervalih shranjujejo podatki iz podatkovnega strežnika. Ker je strežnik večino časa izklopljen, je možnost vdora vanj minimalna.

1. Spletni strežnik teče na virtualnem privatnem strežniku (VPS), kar omogoča enostavno skalabilnost (kloniranje slike sistema na več strežnikov) ter zanesljivost. Promet se do strežnika usmerja s tako imenovanim upravljalnikom obremenitve (ang. load balancer), ki v primeru uporabe več strežnikov promet do njih enakomerno porazdeli, tako da so vsi strežniki enako obremenjeni. Strežnik je od zunaj dostopen (viden) le preko vrat (ang. port) 443, kar nudi dodatno zaščito pred nepooblaščenimi vdori.
2. Podatki se shranjujejo na samostojnem podatkovnem strežniku, ki se nahaja v zavarovanih prostorih znotraj IJS. Zaradi povečane varnosti je dostopen le znotraj lokalne mreže IJS (intranet).
3. Strežnik za varnostne kopije je tretji strežniški sistem uporabljen v projektu, deluje po načelu strežnika NAS (ang. network-attached storage), je pa večino časa izklopljen (ang. offline). Zbudi se v določenih intervalih, ko naredi ali posodobi varnostno kopijo iz podatkov na podatkovnem strežniku.

Komunikacija med vsemi tremi strežniškimi komponentami je zavarovana z dvostranskim šifriranjem (ang. encryption), dostop do njih pa omejen na eno osebo ter lokalno mrežo IJS.

Strežniška shema je prikazana na sliki 2

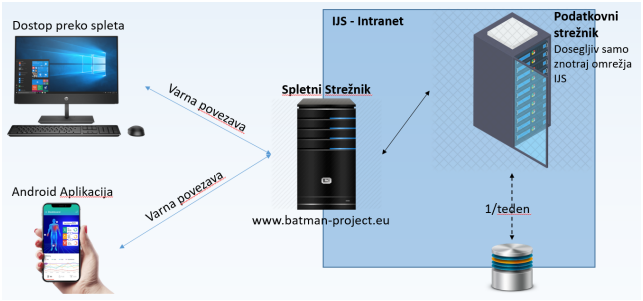

Figure 1: Strežniška Arhitektura v povezavi z dostopnostjo iz intro/outro-omrežja

2.2 Programska rešitev

Projekt BATMAN je v osnovi raziskovalni projekt. To pomeni, da se podatki pacientov zbirajo tudi za namen nadaljnjih raziskav, s katerim se bo izboljšalo razumevanje ter zdravljenje bolezni. Pomembno je, da je sistem zasnovan tako, da hkrati nudi popolno zaščito podatkov pacientov (uporabnikov) ter njihovo anonimnost, hkrati pa raziskovalcem (zdravnikom) omogoči dostop do podatkov, ki so potrebni za raziskave. V projektu ta problem rešujemo s tako imenovanimi *krogi*. Krogi so v grobem skupine uporabnikov, ki si delijo iste pravice (ang. permissions) in prepovedi (ang. prohibitions), pri izvajanju določenih akcij (npr. urejanje nastavitev uporabnika) ter dostopa do informacij. Pravica torej uporabniku dovoljuje, da vidi podatke o DNA uporabnika Y, prepoveduje pa mu ogled fizične aktivnosti.

2.2.1 Varovanje osebnih podatkov

Vsi podatki na strežniku so zaščiteni (šifrirani) na nivoju sistemske slike (ang. system image/partition). To zagotavlja varnost podatkov v primeru nedovoljenega, fizičnega, vdora oz. kraje podatkovnih diskov iz strežnika, ne rešuje pa problema vdora v sistem, ko je ta zagnan, saj so v tem primeru podatki dešifrirani.

Različni podatki imajo različne stopnje pomembnosti in potrebno se je zavedati prednosti (zaščita podatkov) ter slabosti (počasno rokovanje s podatki) šifriranja. Ker nekateri podatki (sami po sebi) kljub morebitnemu razkritju ne ogrožajo anonimnosti oz. zasebnosti uporabnika, jih nima smisla dodatno šifrirati. Tu gre predvsem za podatke senzorjev (npr. podatki iz pospeškomera, srčni utrip, količina gibanja, ipd.), ki sami zase ne povejo ničesar, kar bi omogočilo identifikacijo ali razkrilo podrobnosti o uporabniku.

Ko govorimo o osebnih podatkih pacienta, kot so ime, priimek, letnica rojstva, e-pošta, pa moramo dodatno poskrbeti za varovanje, saj lahko katerikoli od teh podatkov identificira pacienta. Enostavna rešitev v tem primeru je, da se osebni podatki o pacientu enostavno ne hranijo. Vsak pacient ob obisku osebnega zdravnika dobi unikatno identifikacijsko številko ter geslo. Vsi osebni podatki o pacientu se hranijo v pisni obliki pri zdravniku. Problem tega pristopa je v tem, da ima zdravnik, ki je zelo obremenjen že z rednim delom, dodatno delo s tem, da hrani ločeno kartoteko pacienta ter spletnega sistema. Prav tako je uporaba (dolgih) identifikacijskih številk ter gesel težavna za uporabnike, kar privede do pogostega pozabljanja teh informacij ter kontaktiranje (osebno, ne po telefonu) zdravstvenega osebja, ki nato ponastavlja gesla in uporabniška imena uporabnika. Številne raziskave [1] kažejo neizvirnost uporabnikov pri izbiri gesel, kar bi bilo v omenjenem primeru, ko ni enostavnega načina za ponastavitev gesla, saj o pacientu nimamo drugih podatkov (e-pošte), le še stopnjevano. Tako bi z uporabo takšnega pristopa dosegli ravno obraten učinek, namesto izboljšanja varnosti ter anonimnosti uporabnikov bi zaradi enostavnih gesel resno ogrozili in izpostavili vse njihove podatke nepooblaščenim osebam, hkrati pa bi povečali količino dela, ki jo opravlja medicinsko osebje. Boljši način je torej shranjevanje teh (tudi. osebnih) podatkov na strežniku ter skrbno omejevanje dostopa do njih.

Varno shranjevanje gesel je že dolgoletna praksa [5], kjer se geslo, pred shranjevanjem v podatkovno bazo šifrira z uporabo razpršilne (ang. hash) funkcij (npr. MD5, SHA256). Ta način hrambe je varen in preizkušen, problem je le, da ga ni mogoče uporabiti pri ostalih podatkih (*opomba: razpršilne funkcije delujejo le v eno smer, iz razpršenega teksta je nemogoče pridobiti original tekst*). Osebne podatke pacienta je torej potrebno zaščititi z uporabo šifrirnih algoritmov (npr. AES-256, RSA), ki poleg šifriranja omogočajo tudi dešifriranje teksta. Njihova slabost v primerjavi z razpršilnimi funkcijami je višja računska zahtevnost ter izmenjava šifrirnih ključev.

Komunikacija z uporabniki Večina komunikacije dandanes poteka v elektronski obliki, preko e-pošte ali drugih spletnih aplikacij. Takšen način komunikacije je zaželen tudi v zdravstvu, ko gre za nenujne primere, saj zdravstvenemu osebju vzame manj časa, ter ga lahko opravijo, ko niso zaposleni s pregledom pacientov. Problem nastane, če želimo ohraniti anonimnost uporabnikov, saj ima večina uporabnikov v svojem e-poštnem naslovu ime ali priimek (*ime.priimek@e-posta.si*. Reševanja tega problema smo se lotili z uporabo strežniškega posrednika pošte, preko katerega poteka vsa komunikacija. Ko uporabnik želi poslati e-pošto določeni osebi, to stori preko spletnega portala ali na za to določen e-poštni naslov (posta@batman-projekt.eu). Strežnik nato iz naslova sporočila (ang. subject) razbere, komu je pošta namenjena ter jo preusmeri na e-poštni naslov prejemnika. Prejemnik nato odgovori na pošto, ki se preko strežnika posreduje uporabniku. Na ta način se ohrani zasebnost uporabnika, hkrati pa to omogoča enostavno implementacijo oz. integracijo s spletnim pogovornim oknom (ang. online chat), ki omogoča pogovore v realnem času.

2.3 Krogi zasebnosti

Z uporabo šifriranja tako strežniških slik (ang. system image) kot samih podatkov v podatkovni bazi lahko privzamemo, da so podatki varno shranjeni in zaščiteni. Naslednji korak je dodeljevanje dostopa do podatkov uporabnika. Če želimo, da je sistem uporaben tako za raziskovalce kot paciente, morajo raziskovalci dobiti določen dostop do podatkov pacientov, hkrati pa ti podatki ne smejo biti takšni, da bi razkrili identiteto pacienta. Reševanje tega problema, smo se lotili z implementacijo tako imenovanih krogov (ang. circles).

Krogi so ime za skupine uporabnikov, ki si delijo pravice in prepovedi, so hierarhično povezani, vsak uporabnik pa lahko ustvarja lastne kroge ter je vključen v več krogov. Najvišji krog je t.i. *administracijski* krog, v katerem so vsi ostali krogi. Krog je namenjen administratorjem, ki dodeljujejo pravice vsem krogom. Shema krogov je predstavljena na sliki 2

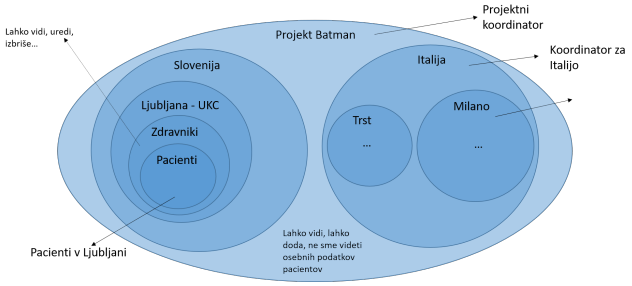

Figure 2: Primer krogov za Slovenijo in Italijo

Sistem je zasnovan tako, da se nove pravice in prepovedi enostavno dodajajo, vsak nižje ležeči krog (potomec) pa lahko vsebuje kvečjemu enake pravice kot njegov starš (ang. parent), saj bi v nasprotnem primeru lahko določeni uporabniki nepravilno dodajali pravice v svojih pod-krogih, tako da bi ti imeli več pravic, kot jih imajo uporabniki sami. Na podoben način se dedujejo prepovedi s to razliko, da potujejo navzgor namesto navzdol. Prepovedi prepovedujejo (izničijo) določeno pravico, ki jo ima nek krog. Pacient dovoljuje dostop do svojih kliničnih podatkov le svojemu zdravniku, ostalim, višje ležečim uporabnikom (koordinator, administrator, ipd.) pa to prepoveduje. Prepoved se bo torej širila od zdravnikovega kroga navzgor. Intuitivno si lahko predstavljamo, da pravice padajo navzdol, medtem ko se prepovedi dvigajo navzgor po hierarhiji (slika 3).

2.4 Mobilna aplikacija - Android

Življenjski stil pacientov znatno vpliva na to, kako se pacient odziva na določene tipe zdravil ter pristopov k zdravljenju, hkrati pa nam do določene mere pove stopnjo bolezni. Kot primer, znano je, da kajenje poslabša simptome bolezni, po drugi strani pa se pacienti zaradi bolečine sorazmerno manj gibljejo. Z uporabo pametnega telefona in njegovih senzorjev lahko dobimo globlji vpogled v pacientov življenjski slog ter avtomatično sklepamo oz. koreliramo določene parametre s stanjem bolezni. Poleg avtomatskega zbiranja podatkov iz senzorjev bo aplikacija uporabniku ponujala možnost izpolnjevanja kratkih anketnih vprašanj o njegovem življenjskem slogu, ki bodo pripomogla k razumevanju njegovih navad v povezavi z boleznijo. V kasnejši fazi projekta, pa bo ponujala odgovore ter predloge k izboljšanju stanja bolezni.

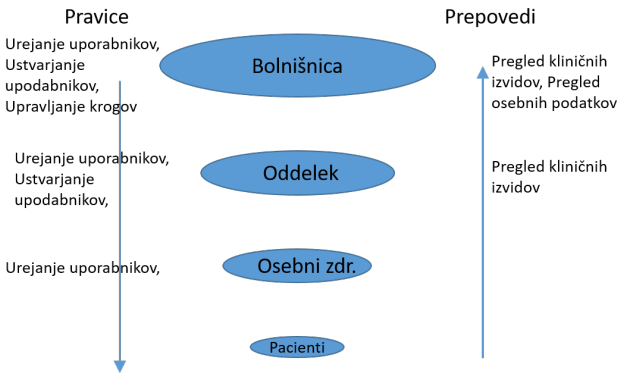

Figure 3: Pravice padajo navzdol, prepovedi pa se dvigajo navzgor

Za vstop v aplikacijo bo potrebna identifikacija z uporabniškim imenom in geslom, na telefonih, ki to podpirajo (imajo strojno opremo za biometrično identifikacijo), pa bo mogoče enostaven prijava z uporabo prstnega odtisa ali druge biometrije (prepoznavanje obraza). *Opomba: Biometrični podatki niso vezani na aplikacijo BATMAN, ampak zanje skrbi operacijski sistem.* Podatki se bodo na telefonu hranili, dokler ne bodo posredovani na strežnik preko za to namenjenih API vmesnikov. Vsa komunikacija s strežnikom bo šifrirana, za identifikacijo pa se bodo uporabljali piškotki s soljo (ang. tokens with salt).

2.5 Kontinuirana integracija in namestitev

Pomembno je, da se vse razvite funkcije redno in večkratno testirajo ob vsaki spremembi kateregakoli dela kode, četudi ta ne vpliva direktno na spremenjeno funkcijo. S tem se doseže zanesljivo delovanje programa ter ustavi morebitne napake, še preden pridejo do uporabnikov. Večkratno testiranje vseh funkcionalnosti, ki jih ponuja spletna stran, je dolgotrajno ter dolgočasno za programerja (kar lahko privede do površnosti), zato jih je potrebno avtomatizirati [3]. V projektu BATMAN s pomočjo avtomatskih *unit*, *funkcijskih* (ang. feature) ter *uporabniških* (ang. browser) testov avtomatsko preverimo, ali vse spisane funkcije delujejo po pričakovanjih, vsakič preden se programska koda naloži na glavni strežnik. To je še posebej pomembno, ko imamo opravka z različnimi pravicami uporabnikov, kjer bi napaka v prikazovanju lahko razkrila osebne podatke pacientov ostalim uporabnikom sistema.

Varnost celotnega sistema temelji na dostopu do šifrirnih ključev ter programske kode, ki z njimi upravlja. Pomembno je, da ima dostop do strežnika le ena oseba, to je glavni administrator za infrastrukturo, razvijalci in ostali, ki razvijajo program, pa ne. To je mogoče storiti z uporabo avtomatske namestitve (ang. automatic deployment [6]), kjer se program po tem, ko je prestal vse avtomatizirane teste ter odobritev za združitev na produkcijski strežnik (ang. merge-request), samodejno prenese na delujoč strežnik. Dodatno se na ta način z uporabo simboličnih povezav ter varnostnih kopij doseže, da je spletna stran dosegljiva vse čas (ang. no downtime).

3. ZAKLJUCEK

V prispevku smo predstavili arhitekturno zasnovo v evropskem projektu BATMAN. V sklopu projekta se bodo zbirali občutljivi medicinski podatki, zato je varnost ter skrb za ohranjanje anonimnosti pacientov na prvem mestu, čemur je podrejena arhitektura ter načrt celotnega programa. Natančneje so v prispevku predstavljeni koraki za zaščito podatkov ter vpeljani t.i. *krogi*, ki skrbijo za dostopnost do podatkov osebam, ki so za to zadolžene.

4. ZAHVALA

Delo je bilo financirano iz projekta BATMAN, iz sklopa projektov ERA PerMed.

5. REFERENCES

[1] J. Bonneau. The science of guessing: analyzing an anonymized corpus of 70 million passwords. In *2012 IEEE Symposium on Security and Privacy*, pages 538–552. IEEE, 2012.

[2] M. Feily, A. Shahrestani, and S. Ramadass. A survey of botnet and botnet detection. In *2009 Third International Conference on Emerging Security Information, Systems and Technologies*, pages 268–273. IEEE, 2009.

[3] M. Fowler and M. Foemmel. Continuous integration. *Thought-Works) http://www. thoughtworks. com/Continuous Integration. pdf*, 122:14, 2006.

[4] J. Isaak and M. J. Hanna. User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, 51(8):56–59, 2018.

[5] J. Katz, A. J. Menezes, P. C. Van Oorschot, and S. A. Vanstone. *Handbook of applied cryptography*. CRC press, 1996.

[6] M. Shahin, M. A. Babar, and L. Zhu. Continuous integration, delivery and deployment: a systematic review on approaches, tools, challenges and practices. *IEEE Access*, 5:3909–3943, 2017.

[7] P. Voigt and A. Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[8] C. C. Zouboulis, N. Desai, L. Emtestam, R. Hunger, D. Ioannides, I. Juhász, J. Lapins, L. Matusiak, E. Prens, J. Revuz, et al. European s1 guideline for the treatment of hidradenitis suppurativa/acne inversa. *Journal of the European Academy of Dermatology and Venereology*, 29(4):619–644, 2015.