OR5 - Trasmissione sicura di dati sensibili

Descrizione sintetica dell'obiettivo

Lo scopo di questo OR è l'analisi di metodologie innovative per la trasmissione di grosse moli di dati provenienti da fonti eterogenee, ponendo particolare attenzione alla gestione della sicurezza e della privacy dei dati sensibili. Le attività previste per il suo svolgimento sono tutte attività di ricerca industriale (RI).

Data la massiccia diffusione di tracker e sensoristica per i segnali vitali e/o medici, quali, ad esempio, pressione sanguigna, battito cardiaco e valore glicemico, nasce la necessità di avere un sistema di trasmissione efficiente e sicuro.

L'aggregazione massiva di dati eterogenei a fini statistici comporta una serie di rischi legati alla privacy della popolazione su cui i dati vengono raccolti. È da ricercare quindi un approccio che permetta ai singoli soggetti il controllo su come i loro dati vengono utilizzati e condivisi, non intaccando al contempo l'utilità degli stessi dati. I problemi che nascono riguardano la confidenzialità, l'integrità, la qualità dei dati e la disponibilità dei dati.

L'OR in esame è stato ottenuto con successo attraverso le attività seguenti.

A5.1 – Tecniche per la trasmissione di grosse moli dati

Obiettivo della Attività

In questa attività saranno analizzate le tecniche avanzate e allo stato dell'arte per il trasferimento di grosse quantità di dati sia di natura testuale, come referti medici, che visiva come immagini biomediche (ecografie, radiografie, etc.). Ne saranno valutati i limiti ed eventualmente si procederà alla definizione di nuove tecniche.

I dati da trasmettere possono essere sia di grande dimensione, sia essere generati in maniera continua e provenire da diverse fonti: basta pensare come esempio all'enorme mole di dati di diagnostica per immagini ad alta risoluzione di tipo DICOM, come tomografia a emissione di positroni (PET), tomografia computerizzata (TC) e risonanza magnetica (RM), che non solo sono per loro natura sono immagini di dimensione non trascurabile, ma vengono anche prodotte da più fonti, generando un elevato traffico. Nasce dunque il bisogno di trovare metodologie efficienti per sincronizzare dati generati dai diversi centri diagnostici.

Descrizione Attività

Nell'attività A5.1 è stato analizzato lo stato dell'arte delle tecniche necessarie a traferire grosse quantità di dati eterogenei. Nonostante i sorprendenti progressi tecnologici compiuti nei settori di Data Analytics e Cloud, permangono alcuni problemi relativi al trasferimento affidabile e alla distribuzione di un grande numero di file e di volumi di Big Data, a velocità sostenute e in maniera distribuita. Tale problema di trasmissione dei Big Data si è inoltre diffuso in tutti i settori, a causa della crescita esponenziale dei dati generati a livello globale. La scienza moderna, ad esempio, è sempre più data-driven e basata su approcci di natura collaborativa. Le simulazioni su larga scala e gli strumenti adottati producono petabyte di dati che vengono successivamente analizzati da decine di migliaia di scienziati presenti in tutto il mondo. Potrebbe sembrare più logico ed efficiente concentrare le risorse di analisi in corrispondenza delle sorgenti dei dati (strumenti o cluster computazionali), ma non sempre questo accade. Al contrario, le soluzioni distribuite (costituite da componenti dislocate geograficamente) sono molto più comuni e su di esse poggiano gran parte delle attività di collaborazione tra i molteplici enti scientifici. Protocolli e strumenti efficienti sono necessari, quindi, per spostare una grande quantità di dati informativi su reti dotate di un'ampia larghezza di banda. In tale attività sono stati presentati dapprima i problemi relativi alla trasmissione di dati in reti WAN. I metodi abituali di trasferimento dei dati (con strumenti basati su scp, http e ftp come curl o wget) funzionano bene quando i dati sono nel range dei MB o GB, ma quando si hanno collezioni di dati molto grandi ci sono alcune considerazioni che è necessario fare, soprattutto se i dati sono trasferiti su Wide Area Networks. È stato indagato il problema della latenza dei pacchetti, sono stati descritti i protocolli di livello trasporto TCP e UDP, e successivamente è stato valutato l'invio in contemporanea di più flussi di pacchetti TCP, come soluzione per il trasferimento dei dati a lunga distanza. Nell'attività A5.1, inoltre, sono descritti i meccanismi alla base del trasporto efficace dei dati, mettendo in evidenza le metodologie utilizzate per la sincronizzazione dei file. Sono poi dettagliate le tecniche per la deduplicazione e sincronizzazione dei dati. La deduplicazione dei dati è il processo di eliminazione delle copie di dati ripetuti, che riduce le ridondanze intra-file e

inter-file, garantendo risparmio di spazio. È stato evidenziato come l'efficacia della deduplicazione dipenda dalla tipologia di algoritmi di deduplicazione e dal set di dati considerati, e come tale operazione fornisca un grande risparmio in termini di spazio occupato, ma risulti "data intensive" e comporti il sovraccarico dell'infrastruttura di storage esistente. Una parte dell'attività ha inoltre riguardato lo studio della compressione dei dati, considerata un caso speciale di data differencing. Successivamente, sono stati passati in rassegna alcuni strumenti per il trasferimento veloce di dati, quali: bbcp, bbftp, lftp, fdt, GridFTP, netcat, tar/netcat e tnc. Sono state inoltre valutate tecniche per la sincronizzazione dei file: quest'ultima è considerata come sottoinsieme della sincronizzazione dei dati. Si tratta di un processo per stabilire una consistenza tra i dati di due storage, sorgente e destinazione, così come la continua armonizzazione dei dati nel tempo. Le tecniche analizzate in dettaglio sono le seguenti: Unison, che divide i file in blocchi disgiunti, confronta questi blocchi e utilizza un algoritmo "rolling hash" per rilevarne le modifiche; Dropbox, sincronizzatore di file "near real-time"; GoodSync, tool per la sincronizzazione bidirezionale e il backup di file tra due dispositivi; Synkron, applicazione multipiattaforma utilizzata per la sincronizzazione di due o più cartelle; DFSR, motore per la replicazione dei dati che permette di mantenere sincronizzate le cartelle su più server, offrendo agli utenti che accedono in remoto un accesso rapido e affidabile ai file, aumentando la disponibilità dei dati; Syncany, che con la tecnica di de-duplicazione tenta di minimizzare il tempo di archiviazione e di sincronizzazione; Rsync, il più noto protocollo single round per la sincronizzazione dei file; Aspera Sync, progettato per superare le carenze prestazionali e di scalabilità degli strumenti di sincronizzazione convenzionali come rsync. Nell'attività A5.1 sono stati inoltre analizzati i file system distribuiti, che supportano la condivisione delle informazioni sotto forma di file e risorse hardware (storage persistente), attraverso una intranet. Le soluzioni studiate sono le seguenti: Sun Network File System (NFS), che segue un modello astratto dove tutte le implementazioni supportano il protocollo NFS; Andrew File System (AFS), che in modo simile a NFS fornisce accesso ai file condivisi remoti per programmi UNIX in esecuzione su workstation; Lustre; Kosmos; Google File System; Panasas; PVFS2; RGFS; MooseFS; iRODS; GlusterFS, file system open source distribuito e scalabile linearmente; HDFS, file system distribuito di Hadoop progettato appositamente per la gestione dei flussi e la memorizzazione affidabile di grandi volumi di dati. Infine, sono state messe a confronto le tecniche per la trasmissione di dati sanitari e le loro caratteristiche. Come descritto in precedenza, la sincronizzazione dei file attraverso reti WAN rappresenta un collo di bottiglia nella trasmissione di dati tra un mittente e un destinatario; pertanto, è stato realizzato uno studio degli algoritmi di trasmissione attraverso una tabella che mostra gli algoritmi e le caratteristiche per ogni tecnica analizzata. Il confronto è basato su: valutazione delle tecnologie di riferimento, velocità con cui sono trasferiti i dati, e le tecniche di crittografia utilizzate per garantire la sicurezza nella trasmissione e nell'archiviazione. Successivamente sono state messe a confronto le funzionalità e le caratteristiche dei file system distribuiti. L'architettura dei DFS deve garantire alte performance di scalabilità all'aumentare delle richieste in ingresso. È presentato un confronto dei DFS in termini di architettura e di funzionalità quali: load balancing, fault detection, replication, naming, gestione delle API. A valle di un confronto prestazionale, infine, è stata proposta la tecnica più vantaggiosa.

A5.2 – Tecniche per la sicurezza delle informazioni

Obiettivo della Attività

L'attività ha come obiettivo quello di valutare le diverse problematiche relative alla sicurezza che concernono il trasferimento di informazioni sensibili contenuti nei dati gestiti dalla piattaforma.

A causa della natura eterogenea delle sorgenti dati utilizzabili, vi sono diversi scenari con diversi livelli di sicurezza da implementare: i) trasmissione di dati clinici, sia testuali che visivi, che necessita di un alto livello di sicurezza; ii) trasmissione di dati aggregati provenienti da sensoristica tra i quali ad esempio, segnali biomedicali e segnali provenienti da fitness tracker per i quali è sufficiente usare una trasmissione con livello di sicurezza "medio".

Saranno studiate le tecniche note in letteratura e definite nuove metodologie per soddisfare i requisiti della trasmissione sicura di dati sensibili.

Descrizione Attività

Nell'ambito della trasmissione di dati provenienti da fonti eterogenee (con diversi livelli di privacy), è necessario prevedere dei meccanismi volti ad assicurare la sicurezza delle informazioni trasmesse. L'attività di ricerca A5.2 ha consentito di valutare diverse problematiche sulla sicurezza, relative soprattutto al trasferimento di dati sensibili dalla piattaforma e verso di essa. Sono state individuate ed analizzate le diverse problematiche relative all'adozione dei meccanismi di sicurezza noti in letteratura per la trasmissione di dati. Sono state quindi analizzate molteplici

metodologie innovative, che consentano di trasmettere grosse moli di dati provenienti da fonti eterogenee, prestando particolare attenzione alla privacy dei dati sensibili. Inoltre, sono stati definiti diversi requisiti di sicurezza da soddisfare nel contesto della trasmissione sicura di dati, riguardanti i seguenti aspetti: autenticazione dei soggetti che prendono parte alla comunicazione; sicurezza del canale, garantendo riservatezza, integrità, autenticità, non-ripudio e disponibilità; autorizzazione nella costruzione del canale, che permette la costruzione del canale di comunicazione in funzione dei diversi livelli di sicurezza associati all'attore che intende utilizzare il canale stesso; autorizzazione nella comunicazione della tipologia di dati. In seguito all'eterogeneità dei dati, sono stati definiti diversi livelli di sicurezza da implementare per la trasmissione sicura delle informazioni da e verso la piattaforma: i) trasmissione di dati clinici, che necessita di un alto livello di sicurezza; ii) trasmissione di dati aggregati provenienti da sensori tra i quali ad esempio, segnali biomedicali e segnali provenienti da fitness tracker, per i quali è sufficiente usare una trasmissione con livello di sicurezza medio. In particolare, è stato svolto uno scouting di soluzioni note in letteratura per la costruzione di canali sicuri nello scambio di informazioni cliniche, mediante lo studio dei diversi protocolli di comunicazione. Inoltre, sono stati individuati i requisiti minimi di sicurezza per la trasmissione sicura di dati clinici in funzione degli attori coinvolti e della tipologia di dati scambiati. L'attività si è quindi concentrata sul trovare soluzioni per effettuare trasferimenti di dati di grosse dimensioni in modo sicuro, e a tal proposito è stata valutata l'adozione di infrastrutture FTP Cloud-based. Tali infrastrutture garantiscono trasferimenti veloci senza alcun limite nella dimensione dei dati trasmessi. Successivamente sono state analizzate tecniche di data integrity, che garantiscano che i dati trasmessi da sorgente a destinazione non vengano alterati. Infatti i dati potrebbero essere intercettati e modificati da utenti malintenzionati, pertanto è opportuno effettuare dei controlli sui dati in transito così che il destinatario sia in grado di confermare che essi non siano stati alterati. Il requisito di data integrity può essere soddisfatto confrontando la fingerprint del file con una base di dati in cui la stessa viene precedentemente memorizzata. La trasmissione sicura di dati sensibili richiede anche un processo di autenticazione; a tal fine, è stato definito un processo di data authentication, che consente di verificare l'autenticità del mittente (e del destinatario) dei dati. Il mittente dovrebbe infatti essere autenticato, in modo tale che un malintenzionato non sia in grado di prendere il suo posto (fingendo di essere il vero mittente), compromettendo l'integrità della trasmissione. È stata infine creata una tabella che descrive le attività mostrate nel processo di costruzione di un canale, per la comunicazione da e verso la piattaforma.

A5.3 – Metodologie per la privacy in ambito medico sanitario

Obiettivo della Attività

Questa attività ha come obiettivo l'analisi e la definizione di metodologie per il rispetto della privacy e per la protezione dei dati medico-sanitari dei soggetti. In particolare, sono stati individuati tre contesti sui quali intervenire andando a definire tecniche innovative: Anonimizzazione, Cifratura, Consapevolezza.

Descrizione Attività

Le novità del Regolamento Europeo sul tema della protezione dei dati personali (GDPR) sono focalizzate sul concetto di sicurezza del trattamento dei dati. Tale regolamento ha modificato radicalmente l'approccio adottato finora per regolamentare la materia, attribuendo un ruolo centrale alla "responsabilizzazione" del titolare e del responsabile del trattamento. Si parla di principio di "accountability", che consiste nell'adozione di modelli organizzativi e appropriate misure tecniche atte a conservare i dati e a gestirli in modo corretto. Pertanto, "tenendo conto dello stato dell'arte e dei costi di attuazione, nonché della natura, dell'oggetto, del contesto e delle finalità del trattamento, come anche del rischio di varia probabilità e gravità per i diritti e le libertà delle persone fisiche", i titolari e i responsabili del trattamento dovranno garantire misure adeguate per fronteggiare i rischi esistenti, adottando tecniche di anonimizzazione, cifratura, pseudonimizzazione e consapevolezza. Nell'attività A5.3 sono state valutate le diverse tecniche esistenti nei contesti summenzionati. Nello specifico, l'anonimizzazione consente la cancellazione, in modo irreversibile, di ogni elemento dal dato personale che consente l'identificazione dell'interessato. Sono stati dettagliati i riferimenti normativi sul tema dell'anonimizzazione e presentati due diversi approcci, basati rispettivamente sulla randomizzazione e sulla generalizzazione. La randomizzazione è un modello basato sull'idea di perturbare i dati da pubblicare aggiungendo una quantità di rumore, e può essere ottenuta tramite: aggiunta di rumore statistico, che consiste nel modificare gli attributi contenuti nell'insieme di dati in modo tale da renderli meno accurati, mantenendo nel contempo la distribuzione generale; permutazione, che consiste nel mescolare casualmente i valori immessi relativi ai dati di trattamento da anonimizzare, permutandoli e disaccoppiandoli; privacy differenziale, che suggerisce al responsabile del trattamento la quantità e la forma di rumore statistico che va aggiunto, per ottenere le garanzie di tutela della sfera privata richieste. Per la generalizzazione, infine, sono state presentate le tecniche di k-anonimity, l-diversità e t-vicinanza. Nell'attività

A5.3, inoltre, è presentata la pseudonimizzazione come processo che blocca la correlabilità dei dati personali all'identità di una persona, ma che non produce un insieme di informazioni anonime, con conseguente possibile re-identificazione del soggetto interessato. Nel testo del GDPR, la pseudonimizzazione è spesso associata ai concetti di "privacy by design" e "privacy by default". È stato poi affrontato il tema della cifratura dei dati. Cifratura e pseudonimizzazione sono tecniche differenti ma con lo stesso scopo: rendere il dato incomprensibile a chi non ha le autorizzazioni necessarie per accedervi. Per la crittografia sono stati dettagliati algoritmi tramite i quali, adottando una opportuna "passphrase", sono "aperti" e "chiusi" i dati. Il GDPR propone l'obiettivo di cifrare server e, più in generali, tutti i sistemi che gestiscono credenziali e trattano i dati sensibili nel settore sanitario, in modo da offuscare tutti i dati presenti in chiaro. Nell'attività A5.3, infine, sono state individuate cinque classi di privacy: Low, Moderate, Normal, Restricted e Very Restricted. A tali classi di privacy sono state assegnate le tecniche di autenticazione adottate per la condivisione di informazioni sulla piattaforma, e la tipologia di dati trattati.

A5.4 – Tecniche per l'accesso e l'autorizzazione

Obiettivo della Attività

L'attività ha come obiettivo l'analisi e la definizione di nuovi sistemi di sicurezza in grado di offrire un meccanismo di autorizzazione per la memorizzazione, da parte di attori eterogenei, sulla piattaforma dei dati e allo stesso tempo garantire l'accesso a diversi livelli di contenuto informativo in funzione al livello di autorizzazione dell'utente. La fase di autorizzazione per l'inserimento e per il recupero di informazioni nella piattaforma si baserà principalmente sulla tipologia di utente che effettua la richiesta. Saranno, infatti, individuate le tipologie di utenti basandosi sui ruoli individuati nel DPCM 29 settembre 2015 (Regolamento in materia di Fascicolo Sanitario Elettronico) n. 178, di cui al comma 7 dell'art. 12 del decreto legge 18 ottobre 2012, n. 179, convertito, con modificazioni, dalla legge 17 dicembre 2012, n. 221, e successive modificazioni. I diversi attori che utilizzeranno la piattaforma dei dati dovranno essere "mappati" adeguatamente sui ruoli definiti: ad esempio sarà possibile mappare i diversi dispositivi di monitoraggio wearable con il ruolo assistito, il quale potrà "scrivere" nella sezione riservata.

Descrizione Attività

Nell'ambito dell'attività A5.4 "Tecniche per l'accesso e l'autorizzazione" del progetto BDA4PHR, sono stati analizzati sistemi di sicurezza per la protezione dei dati sanitari memorizzati in sistemi informativi. Nello specifico, sono stati analizzati e valutati diversi meccanismi di controllo degli accessi noti in letteratura. Il controllo degli accessi è uno dei meccanismi di sicurezza più utilizzato nei sistemi informativi, che permette il soddisfacimento dei requisiti di integrità, confidenzialità e qualità dei dati. Un meccanismo di controllo degli accessi permette di limitare l'accesso ai documenti, ai dati e alle funzionalità del sistema, consentendo di specificare chi può operare e come può farlo.

Alcuni dei modelli di controllo degli accessi più importanti in letteratura e dettagliati nell'attività A5.4 sono i seguenti: i) Mandatory Access Control (MAC), che si basa sulla specifica a priori di una serie di attributi di sicurezza per soggetti (utenti fruitori delle risorse informative) e oggetti (le risorse informative fruite). Quando un soggetto tenta l'accesso ad un oggetto informativo, il controllo degli accessi esamina gli attributi di sicurezza del soggetto richiedente e dell'oggetto richiesto. In funzione di questi due attributi di sicurezza, decide se l'accesso può essere consentito oppure deve essere negato; ii) Discretionary Access Control (DAC), in cui il proprietario delle risorse informative ha la facoltà di concedere l'accesso attraverso la definizione di una specifica politica di sicurezza; iii) Role-Based Access Control (RBAC), in cui l'elemento fondamentale è il ruolo. Ad ogni utente del sistema sarà associato un ruolo, in funzione del quale l'utente potrà effettuare una serie di operazioni sulle risorse; iv) Attribute-Based Access Control (ABAC), che garantisce l'accesso alle risorse da parte degli utenti attraverso l'utilizzo di politiche di accesso basate su attributi; v) Privacy Centric Access Control, orientato al soddisfacimento del requisito di privacy dei dati sensibili. Nell'attività A5.4, inoltre, è stata inserita una tabella sinottica comparativa dei modelli di AC, basata sulle seguenti caratteristiche: flessibilità, granularità, dinamicità, gestione delle politiche di accesso, semplicità computazionale. Successivamente, è stata presentata una tabella in cui, per ogni macro-categoria individuata per l'accesso ai dati gestiti dalla piattaforma di PHR, è presente una descrizione e la corrispondenza del ruolo indicato nel DPCM. Infine, l'attività A5.4 si chiude con la descrizione di un modello di comunicazione sicuro per lo scambio di attributi su piattaforma e tra sistemi federati.