OR2 - Archiviazione, indicizzazione e reperimento di informazioni sanitarie e personali

Descrizione sintetica dell'obiettivo

Definire metodologie innovative per l'immagazzinamento di grosse moli di dati di tipo clinico-sanitario attraverso infrastrutture distribuite di Cloud Computing e analisi delle più recenti tecniche di indicizzazione, reperimento e strutturazione automatica di documenti archiviati su Cloud.

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

A2.1 – Analisi di tecniche di Data Storage e Clouding

Obiettivo della Attività

Analisi delle metodologie di Cloud Computing più adatte a supportare il sistema di data storage della piattaforma.

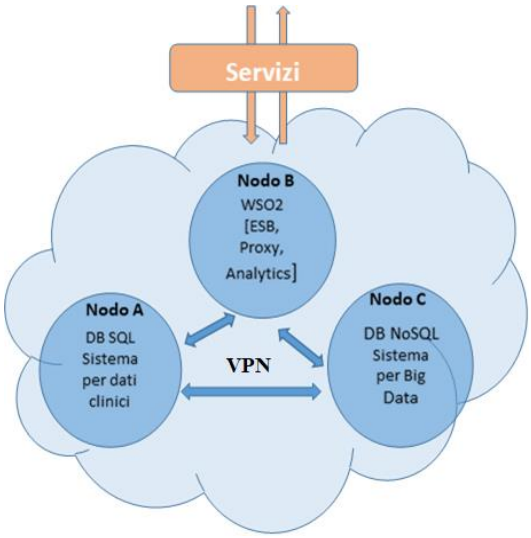
Descrizione Attività

Nell'attività sono state analizzate le migliori configurazione cloud secondo la letteratura corrente e lo stack tecnologico più idoneo alla soluzione che si andrà ad implementare all'interno del progetto.

Sono stati definiti dei modelli concettuali di storage riguardanti la parte relazionale per i dati clinici, producendo i diagrammi E-R delle aree Paziente, Attività, Azioni, Risorse, Dati Clinici e dati amministrativi. Inoltre sono state approfonditi i protocolli HL7 (Health Level Seven) e gli standard ISO, CEN e UNI/UNIFO.

Per la parte riguardante il cloud, invece, sono stati verificati i cloud di tipo NIST per le entità Consumer, Provider, Auditor, Broker e Carrier; inoltre sono state approfondite le problematiche dei modelli di distribuzione di tipo Private, public e Hybrid Cloud.

Sono stati analizzati e descritti le funzionalità dei cloud forniti dal mercato, prendendo in esame i cloud computing offerti da Google, Amazon e Aruba.



A2.2 – Definizione di tecniche di Big Data Clouding

Obiettivo della Attività

Definizione di tecniche innovative di data storage e data repository dedicate a documenti, dati ed immagini di tipi clinico/sanitario e adatte ad essere successivamente utilizzate su infrastruttura di Cloud Computing.

Descrizione Attività

La prima parte dell'attività ha riguardato l'analisi delle sorgenti dalle quali verranno reperiti i dati e analizzati uno ad uno, così da definire alcune tecniche di storage utili al progetto, orientati al paradigma dei Big Data. I dati che si andranno ad analizzare saranno di diversa natura: immagini ad alta definizione, pdf, testo, valori numerici, etc.

Per la parte dei Big Data sono state analizzate le 3V caratteristiche dei paradigmi di Big Data (Volume, Varierà e Velocità), aggiungendo anche quelli di Valore e Veridicità.

Inoltre sono stati analizzato gli aspetti di gestione riguardo alla qualità dei dati e alla loro manipolazione e governance.

Sono state prese in considerazione alcune metodologie esistenti per il Big Data Management (HDFS, MapReduce, ZooKeeper, HBase, Hive e Pig).

Infine sono state esaminate alcune soluzioni tra cui Redis e MongoDB. MongoDB è stato scelto per la sua flessibilità e capacità di gestire documenti e RedIs è stato valutato come possibile candidato per implementare la sezione di caching dei dati clinici.

A2.3 – Definizione di tecniche innovative per l'indicizzazione e la strutturazione di documenti

Obiettivo della Attività

Identificare entità mediche rilevanti in maniera automatica all'interno di testi narrativi (Diario clinico, referti, diagnosi, file HL7, CDA2, ecc).

Descrizione Attività

L'attività A2.3, attraverso l'interazione con ICAR-CNR, ha visto lo studio dei primi algoritmi NLP presenti in letteratura scientifica e sul mercato; sono state analizzate le metodologie più adatte agli scopi del progetto, prestando particolare attenzione sia alle modalità di definizione degli annotatori sia alla individuazione dei vocabolari controllati clinici da includere: inoltre, in accordo con i tipi di dati rilevati nell'attività 2.2, si sono tracciati i criteri di indicizzazione per ogni tipologia di dato provenienti da fonti esterne (sensori, direttamente dal paziente, da altre fonti strutturati, ecc.).

Per questa attività sono state considerate diverse tipologie di documenti come sorgenti di dati da cui estrarre le entità mediche di interesse: testo narrativo inserito dall'utente, documenti clinici in linguaggio naturale e file di tipo strutturato, contenenti parti di testo libero (come ad esempio HL7 CDA2, HL7 FHIR).

Per individuare specifici tipi di informazione all'interno dei documenti testuali, sono state utilizzate tecniche di Information Extraction (IE), Medical Entity Recognition e Relation Extraction, appositamente sviluppate per trattare la complessità del linguaggio medico e le particolarità della lingua italiana. In particolare sono state applicate in fase di analisi tecniche di NLP in grado di effettuare le seguenti operazioni: sentence detection, word tokenization, part-of-speech tagging, dictionary look-up annotation e parsing rules matching. L'output prodotto (atteso) è costituito da documenti indicizzati e strutturati, al fine di poter reperire rapidamente informazioni, effettuare ricerche semantiche e mostrare in maniera sintetica la storia clinica di un paziente. La definizione di tali tecniche di indexing permetterà di usufruire dei servizi offerti dalla piattaforma e di agevolare l'Information Retrieval (IR) dei dati clinici di interesse mediante la rappresentazione degli stessi utilizzando standard di rappresentazione delle informazioni sanitarie come FHIR.