POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Informatica

Tesi di Laurea Magistrale

Automatizzazione dei controlli di sicurezza per la gestione del ciclo di vita dei dati

Relatore
Prof. Fulvio Valenza

Candidato
Ilaria D'Ancona

Anno Accademico 2022-2023

Sommario

In uno scenario globale caratterizzato da un'evoluzione tecnologica e digitale senza precedenti, i dati sono considerati una risorsa dal valore inestimabile, la cui protezione dovrebbe rappresentare una priorità. Tuttavia, garantire la sicurezza delle informazioni rappresenta oggi una sfida ardua a causa dell'enorme quantità di dati generati, dell'aumento di minacce informatiche e del crescente utilizzo di applicazioni basate sul cloud, dove i dati vengono archiviati e condivisi in un ambiente privo di perimetri definiti.

Una violazione dei dati può comprometterne la riservatezza, l'integrità o la disponibilità, con conseguenze significative sia per le aziende che per i singoli individui. Danni reputazionali, perdite economiche, furti d'identità, sono alcuni dei rischi in cui si può incorrere in assenza di un'adeguata protezione delle informazioni, in aggiunta a severe sanzioni in caso di non conformità alle normative vigenti che regolamentano il trattamento dei dati sensibili.

In questa prospettiva, per le imprese è fondamentale sviluppare una strategia di data governance: gestione efficace ed efficiente delle informazioni attraverso l'uso di tecnologie all'avanguardia e misure di sicurezza per la difesa da attacchi informatici, da fughe accidentali di dati o da uso scorretto dei sistemi informatici.

In alcuni settori come quello finanziario, dove vengono trattati per lo più dati sensibili, affidarsi ad una strategia efficiente di data governance può fare la differenza. Pertanto, il lavoro presentato in questa tesi ha lo scopo di proporre una soluzione di data governance aziendale con particolare riferimento all'ambiente bancario, ma con potenziale applicazione in diversi settori.

L'elaborato descrive una pipeline di azioni e controlli di sicurezza automatizzati adottati per proteggere i dati durante il loro ciclo di vita, dalla creazione alla cancellazione.

Il lavoro è stato avviato impiegando un sistema di classificazione automatica basato sul livello di riservatezza di dati non strutturati, ospitati in un ambiente di collaborazione e archiviazione su cloud. I permessi di accesso alla piattaforma cloud in questione sono stati monitorati realizzando appositamente un automatismo. Sulle base delle vulnerabilità emerse dalle due attività precedenti, è stata svolta un'analisi qualitativa del rischio informatico. Infine, si è stabilita una politica, in conformità con i requisiti normativi, per la definizione e l'applicazione di un periodo minimo di conservazione dei dati in base alla loro tipologia.

I risultati ottenuti dimostrano come, attraverso l'uso di automatismi, sia possibile ottimizzare tanto il processo di classificazione, quanto quello di identificazione delle potenziali situazioni di rischio nella gestione degli accessi.

Con questo lavoro si sono volute proporre soluzioni pratiche per una gestione responsabile dei dati informatici tramite controlli di sicurezza automatizzati e possibili azioni

correttive. Queste soluzioni rappresentano un primo passo verso l'automazione della sicurezza dei dati, pur tenendo conto dei limiti legati agli strumenti disponibili, alla sensibilità dei dati trattati e alla necessità di combinare l'automazione con il controllo umano.

Ringraziamenti

Ci tengo a dedicare questo spazio alle persone che, con il loro supporto, mi hanno sostenuto nel percorso per raggiungere questo traguardo.

Vorrei ringraziare il mio relatore, il professore Fulvio Valenza, per avermi fornito spunti fondamentali per la stesura di questo lavoro.
Un sentito ringraziamento va al mio tutor Nicola, che mi ha seguito, con disponibilità e gentilezza, durante il tirocinio e nei vari passi della realizzazione di questo elaborato.

Ringrazio infinitamente la mia famiglia per avermi sempre sostenuta, appoggiando ogni mia decisione e permettendomi di affrontare serenamente tutto il mio percorso di studi.
Un ringraziamento speciale lo dedico ad Alberto, per aver sempre creduto in me ed essermi stato accanto con pazienza.
Un ultimo ringraziamento va a tutti i miei amici e ai miei compagni di università, per avermi accompagnato in questo lungo percorso universitario.

Indice

Elenco delle tabelle

Elenco delle figure

Capitolo 1

Introduzione

Nell'era della trasformazione digitale, ogni giorno, miliardi di persone generano, trasmettono, elaborano e archiviano un' enorme quantità e varietà di informazioni attraverso l'uso di dispositivi sempre più sofisticati. Per le organizzazioni di ogni settore questi dati rappresentano una risorsa preziosa e possono essere sfruttati per stimolare l'innovazione e ottenere un vantaggio competitivo.

Per far fronte all'incremento dei dati raccolti, negli ultimi anni si è assistito a un aumento di soluzioni cloud, che offrono non solo ampi spazi di archiviazione, ma anche una varietà di strumenti e nuove tecnologie facilmente accessibili senza la necessità di investimenti significativi.

La gestione di enormi volumi di informazioni, tuttavia, porta con sé la sfida di garantire la loro sicurezza da potenziali rischi, come attacchi informatici, perdite, furti o errori umani. Una violazione dei dati può compromettere la riservatezza, l'integrità e la disponibilità delle informazioni, con ripercussioni significative per le aziende e per i singoli individui, come danni alla reputazione, perdite economiche, furti d'identità, violazioni della privacy. La protezione delle informazioni, infatti, non è solo una questione tecnica, ma coinvolge anche aspetti legali ed etici. È necessario rispettare normative sempre più rigorose e complesse che salvaguardano i diritti e le libertà delle persone in relazione al trattamento dei loro dati personali

I requisiti necessari per soddisfare tutte le esigenze di sicurezza risultano oggi sempre più complessi e interconnessi, rendendo impegnativa la conciliazione tra le funzioni aziendali, la protezione dei dati e la conformità normativa. Le aziende devono affrontare la questione su diversi fronti, accelerando il loro percorso verso una maggiore sicurezza.

La presente tesi di laurea nasce come risultato di un'esperienza di tirocinio in azienda e propone l'introduzione di controlli automatizzati per la protezione dei dati, con particolare riferimento al caso di studio dell'azienda stessa.

L'approccio adottato affronta la questione della sicurezza dei dati seguendo il loro ciclo di vita, con l'intento di migliorare la protezione delle informazioni attraverso il supporto

all'utente nella gestione dei documenti riservati, il monitoraggio dei permessi di accesso alle informazioni e la definizione delle modalità di conservazione dei dati.

L'elaborato descrive le varie fasi del progetto e si suddivide nei seguenti capitoli e relativi contenuti:

- **Capitolo 2**: introduzione dei principali concetti teorici utili a comprendere le problematiche affrontate nell'elaborato.
- **Capitolo 3**: discussione degli obiettivi dettagliati che hanno guidato ciascuna delle fasi del lavoro.
- **Capitolo 4**: descrizione del flusso di lavoro e breve introduzione degli strumenti utilizzati.
- **Capitolo 5**: spiegazione dettagliata delle strategie di sicurezza implementate.
- **Capitolo 6**: definizione del modello di rischio e analisi.
- **Capitolo 7**: discussione dei risultati ottenuti dall'implementazione delle strategie proposte, rivalutazione del rischio e proposte di sviluppi futuri.
- **Capitolo 8**: conclusioni sul lavoro svolto.

Capitolo 2

Fondamenti Teorici

Questo capitolo presenta i principali concetti teorici utili a comprendere le problematiche affrontate nell'elaborato.

L'obiettivo della tesi è stato quello di integrare soluzioni automatiche nei controlli di sicurezza dei dati nel corso del loro ciclo di vita sull'infrastruttura Cloud, includendo la classificazione appropriata, la gestione dei permessi di accesso e il rispetto della conservazione minima richiesta per legge. Il tutto è stato accompagnato da un'analisi del rischio informatico associato alla gestione dei dati aziendali.

Il capitolo inizia con una definizione del concetto di "dato" e delle sue caratteristiche. Viene esaminato il quadro legislativo vigente in materia di protezione dei dati, sia a livello nazionale che internazionale. Viene poi descritto il concetto di rischio informatico e le procedure di valutazione dello stesso. Infine, viene introdotto il Cloud Computing, per offrire una panoramica dell'ambiente in cui si è svolto il lavoro, illustrando le caratteristiche, i benefici e le sfide.

2.1 I dati

Il concetto di "Dato", inteso come elemento di informazione costituita da simboli, ha un'origine relativamente recente, intorno agli anni 60 del secolo scorso con l'avvento dell'informatica moderna. In termini semplici, un dato può essere considerato come un singolo fatto o una descrizione di qualcosa, spesso rappresentato sotto forma di numeri, testo, immagini o altri formati digitali.

Con il passare degli anni, l'espansione delle tecnologie digitali ha innescato un notevole aumento nella quantità di dati generati e disponibili. Ogni giorno nei contesti aziendali si creano centinaia di informazioni, che assumono un significato specifico quando contestualizzati all'interno di un processo aziendale.

Attualmente, le principali fonti di dati sono rappresentate dal Web2.0 [18], che comprende una vasta gamma di piattaforme online interattive e collaborative, e dall'Internet

delle Cose (Internet of Things, "IoT"). Quest'ultima rappresenta un'importante fonte di dati, poiché dispositivi connessi rilevano e trasmettono informazioni in tempo reale, contribuendo a generare un quadro più completo e dettagliato del mondo digitale.

La crescente quantità di dati, risulta di scarsa utilità senza la disponibilità di spazi di archiviazione adeguati. Come sostenuto da Loukides [18], la tendenza dei dati è quella di espandersi fino a saturare completamente gli spazi a disposizione. Di conseguenza, nel corso del tempo, è diventato imperativo ampliare continuamente le capacità di archiviazione per far fronte alla crescita incessante del volume di dati. A questo scopo sono nati i Cloud, piattaforme online che offrono grandi spazi di archiviazione e che verranno introdotti nella Sezione 2.4. La maggior parte dei dati, però, non è destinata ad essere conservata per sempre ed è essenziale implementare strategie efficaci di gestione del ciclo di vita dei dati che ne definiscano le fasi chiave, dalla creazione all'eliminazione. Una gestione oculata del ciclo di vita del dato, non solo ottimizza l'uso degli spazi di archiviazione, ma contribuisce anche a garantire la conformità normativa e a preservare i dati di valore per il periodo appropriato.

I dati sono considerati come il "nuovo oro", evidenziando la loro straordinaria valenza economica. Le aziende sono sempre più orientate a collezionare quantità massicce di dati perché ciò consente loro di estrarre informazioni significative, prendere decisioni informate e migliorare le proprie operazioni.

Questa crescente importanza dei dati ha portato alla nascita di discipline specifiche, come la "Data Science", che si occupa dello studio avanzato dei dati per ottenere deduzioni approfondite e previsioni. Gli scienziati dei dati utilizzano algoritmi, modelli statistici e strumenti analitici per analizzare e interpretare i dati, contribuendo così a guidare decisioni informate nelle organizzazioni.

Tuttavia, questa massiccia raccolta e utilizzo dei dati ha anche sollevato questioni etiche relative alla privacy. Di conseguenza, sono diventate necessarie regolamentazioni rigorose per garantire una gestione responsabile. La necessità di bilanciare l'innovazione e l'utilità dei dati con la tutela della privacy individuale è diventata una sfida chiave nel contesto di questa rapida evoluzione digitale.

2.1.1 Ciclo di vita dei dati

Il concetto di ciclo di vita del dato si riferisce all'insieme delle fasi che un dato percorre dalla sua nascita alla sua eliminazione. Queste fasi comprendono la creazione, la memorizzazione, l'utilizzo, la condivisione, la trasmissione, la modifica, l'archiviazione e la distruzione dei dati. In altre parole, il ciclo di vita del dato determina il periodo di tempo in cui un dato esiste e può essere trattato.

A seconda della tipologia e della natura dei dati, il ciclo di vita può variare e richiedere una gestione adeguata. In ogni fase del ciclo di vita del dato, è fondamentale stabilire chi ha la responsabilità e l'autorità di gestire i dati, quali sono le attività da svolgere e quali

sono le misure di sicurezza da adottare. Questo serve a garantire la protezione dei dati da possibili minacce, come accessi illeciti, alterazioni non autorizzate o perdite accidentali.

A questo scopo, è necessario definire una valida strategia di gestione del ciclo di vita del dato per tenere conto del livello di sensibilità e dei requisiti legali e aziendali relativi ai dati. Tra gli aspetti principali che devono essere tenuti in considerazione ci sono: la durata di conservazione dei dati, i processi di distruzione, la gestione degli accessi, la protezione e la condivisione.

Per enfatizzare ulteriormente il concetto di importanza della gestione del ciclo di vita delle informazioni, si è svolta una ricerca su questo argomento che ha portato alla scoperta di diversi studi che esaminano i dati con un approccio orientato alla loro protezione durante l'intero ciclo di vita. Per esempio, nell'articolo di Yu and Wen [35], gli autori si concentrano sui dati immagazzinati su Cloud e sostengono che, per mitigare il problema della sicurezza, bisogna partire proprio dal ciclo di vita del dato. Quindi, viene suggerito un modello di progettazione per le misure di sicurezza che tiene conto delle singole fasi del ciclo di vita e delle minacce che possono comprometterne l'integrità durante ognuna di queste. Anche nello studio di Koo et al. [17] viene affrontato il problema relativo alla sicurezza dei dati, ampliando l'analisi ai "Big Data" in generale. In questo lavoro sono state identificate le minacce e i problemi di sicurezza che si verificano nel ciclo di vita dei "Big Data", confrontando le linee guida sviluppate dalle organizzazioni internazionali di standardizzazione e analizzando gli studi correlati. Inoltre, gli autori hanno suddiviso il ciclo di vita in cinque fasi e definito una tassonomia in base alle minacce e ai problemi di sicurezza identificati. Gli autori Zhang et al. [36] hanno discusso l'importanza della protezione della privacy durante ciascuna delle fasi del ciclo di vita, soffermandosi sui dati riguardanti le informazioni personali, dalla creazione da parte degli individui fino alla distruzione dopo l'uso per lo scopo previsto.

La diffusione di ricerche che si interessano all'analisi dei dati considerandone l'intero ciclo di vita, evidenzia come questo sia un approccio esaustivo e applicabile in vari contesti.

Il ciclo di vita del dato può essere suddiviso in cinque fasi principali, come illustrato nella Figura 2.1. Di seguito viene riportato un elenco contenente una descrizione per

Figura 2.1. Fasi del ciclo di vita del dato

ciascuna delle cinque fasi:

1. **Creazione**: in questa fase, il dato nasce o viene raccolto da una sorgente che può appartenere o meno all'organizzazione. I dati possono essere prodotti sia da persone che interagiscono con il Web, con applicazioni online o con altri dispositivi, sia da macchine o sensori che misurano vari fenomeni fisici o ambientali. Una volta creato, un dato entra a far parte dei processi aziendali che lo impiegano per vari scopi.

2. **Conservazione**: in questa fase, il dato viene memorizzato in appositi spazi fisici o virtuali, come database, cloud, ecc. Il dato viene organizzato e strutturato attraverso l'uso di indici, in modo da facilitarne il recupero e l'utilizzo successivo.

3. **Condivisione e utilizzo**: in questa fase, il dato viene trattato e sfruttato per alimentare i processi aziendali. Queste attività comprendono la sua analisi, trasformazione, visualizzazione e condivisione. Il dato può fornire informazioni utili per il supporto alle decisioni, la creazione di valore, l'innovazione, ecc.

4. **Archiviazione**: in questa fase, il dato viene nuovamente memorizzato in attesa di essere eliminato o riutilizzato. Inoltre, può essere spostato in altri spazi di memorizzazione, come backup e archivi. Il dato può essere soggetto a revisioni periodiche per verificarne la validità e rilevanza.

5. **Cancellazione**: in questa fase, il dato viene cancellato. Questo può avvenire per motivi di sicurezza, privacy, conformità e obsolescenza. Il dato viene eliminato definitivamente, in modo da non poter essere più letto oppure associato a una fonte o a un soggetto identificabile.

Considerata la varietà dei dati che è possibile trovare, un primo passo per una gestione efficiente consiste nel classificare i dati al momento della loro creazione o del loro inserimento nei sistemi aziendali. La classificazione, che consiste nell'assegnare ai dati una categoria in base alla loro natura, origine, destinazione e livello di riservatezza, permette di identificare quelli più importanti o sensibili, quindi di adottare le misure di protezione più adeguate. Una volta individuata la categoria, è possibile ottenere una prima indicazione del ciclo di vita di quel dato.

Durante la fase di memorizzazione, devono essere scelti il supporto, il formato e la modalità di archiviazione più appropriati. Vanno tenuti in considerazione i requisiti di sicurezza, accessibilità, integrità e durabilità dei dati e vanno applicati eventuali meccanismi di protezione, come la cifratura, per i dati riservati o sensibili.

Durante la fase di utilizzo e condivisione dei dati, è fondamentale stabilire regole per l'accesso, la modifica e la condivisione dei dati. È possibile adottare misure preventive per limitare eventuali violazioni, abusi o perdite di dati scegliendo metodi di condivisione delle informazioni appropriati e monitorando costantemente gli accessi.

Per quanto riguarda l'archiviazione, bisogna prestare attenzione nel mantenere i dati in uno stato di conservazione adeguato per il tempo necessario al raggiungimento delle

finalità per cui sono stati raccolti. In questa fase, si devono osservare i termini e le modalità di conservazione previsti dalle normative applicabili, e si devono garantire le condizioni di sicurezza, qualità e aggiornamento dei dati.

L'ultima fase, quella di cancellazione, richiede il rispetto di determinati vincoli prima di poter procedere alla distruzione definitiva dei dati. É necessario assicurarsi che le informazioni non vengano eliminate se sono soggette a vincoli legali che ne richiedono la conservazione per un periodo minimo non ancora terminato. Al contrario, è importante assicurarsi che le informazioni che per legge devono essere mantenute nei sistemi per il minor tempo possibile, siano tempestivamente eliminate quando non più utili per le finalità per cui sono state raccolte. In questa fase è importante adottare delle tecniche di distruzione efficaci per impedire il recupero dei dati.

2.1.2 Tipologie di dati

Quando si parla di sicurezza dei dati non si può prescindere dal parlare di tipologie di dato. Infatti, questi, si possono suddividere principalmente in tre tipologie: dati strutturati, dati non strutturati e dati semi-strutturati. Ciascuna è caratterizzata da elementi differenti, pertanto è necessario discernere il trattamento e la gestione degli stessi sulla base delle informazioni distintive.

Entrando più nel dettaglio, i dati strutturati vengono così definiti per la natura predefinita dei loro formati. Tipicamente, i dati appartenenti a questa tipologia sono composti da numeri, date e stringhe raggruppati in tabelle o basi di dati e possono essere visualizzati in modo intuitivo attraverso righe e colonne. Avendo degli schemi predefiniti, permettono una migliore gestione degli spazi di archiviazione che si traduce in minor spazio occupato e maggiore velocità di lavorazione. Inoltre, i formati con i quali vengono rappresentati questi dati sono tali da non poter essere modificati. Perciò, le aziende possono sfruttare tecniche di progettazione e di protezione disponibili nel mercato e vagliati dai maggiori produttori, con la garanzia che anche in futuro tali implementazioni rimangano valide.

Questa tipologia di dati, solitamente, viene immagazzinata in basi di dati relazionali (RDBMS), ovvero in raccolte di informazioni che organizzano i dati in relazioni predefinite, alle quali è possibile accedere con delle richieste effettuate tramite il linguaggio di programmazione SQL (Structured Query Language). Inoltre, l'immagazzinamento e la gestione di questi dati seguono l'approccio "Schema-on-write", vale a dire che lo schema del dato viene definito prima che il dato venga inserito all'interno della base di dati. Questo approccio e questa configurazione permettono di rendere i dati ben organizzati e semplici da lavorare per i linguaggi macchina, velocizzando operazioni come inserimento, ricerca e manipolazione.

I vantaggi e gli svantaggi relativi all'utilizzo di dati strutturati possono essere raggruppati ed elencati come segue:

- **Vantaggi**:

  - Sono più semplici da gestire.
  - I tempi di lavorazione sono ridotti.
  - Possono essere sfruttati più facilmente dagli algoritmi di machine learning.
  - Esistono diversi strumenti disponibili per l'utilizzo, l'analisi e la gestione perchè questa categoria esiste da parecchi anni.

- **Svantaggi**:
  - La loro struttura predefinita ne limita l'utilizzo a casi d'uso specifici.
  - Le basi di dati dove vengono immagazzinati sono poco flessibili.

Per quanto riguarda i dati non strutturati, a differenza della categoria descritta sopra, questi non seguono dei formati predefiniti. Infatti, questa categoria racchiude tutto un insieme di tipologie diverse di dati che, una volta immagazzinati, mantengono i loro formati originali, tra cui testo, immagini, audio e video. L'assenza di uno schema unico rende l'analisi, l'organizzazione e la protezione di questi dati più complessa e richiede uno spazio di archiviazione maggiore rispetto a quelli strutturati. Secondo diversi studi, come quello condotto da Gemson Andrew Ebenezer and Durga [10], si stima che ad oggi la percentuale di dati non strutturati in circolazione raggiunge circa il 90% del totale e che soltanto il 10% è composto da dati strutturati. Inoltre, le previsioni sulla distribuzione di queste categorie nei prossimi anni mostrano che queste percentuali sono destinate a crescere in favore dei dati non strutturati.

Tipicamente questi dati sono immagazzinati nei così detti "data lakes", ovvero degli archivi centralizzati progettati per archiviare e proteggere grandi quantità di dati non strutturati (ma anche strutturati e semi-strutturati) nel loro formato nativo e di elaborarne qualsiasi varietà, ignorando i limiti di dimensione. L'approccio più comune per immagazzinare e gestire questa categoria viene definito "Schema-on-read" e consiste nel assegnare uno schema al dato nel momento in cui viene letto a seconda delle esigenze richieste.

I casi d'uso per i dati non strutturati sono svariati. Tra i principali troviamo l'allenamento di modelli di deep learning per il riconoscimento e la classificazione automatici, ad esempio per le immagini e i suoni. O ancora, la lavorazione e la comprensione del linguaggio naturale da parte delle macchine per effettuare analisi del testo o per costruire modelli generativi di testo, come i "Large Language Models" (LLMs) che negli ultimi anni hanno trovato ampio impiego in svariati campi (es. ChatGPT).

I vantaggi e gli svantaggi relativi all'utilizzo di dati strutturati possono essere raggruppati ed elencati come segue:

- **Vantaggi**:
  - Un vasto insieme di formati facilita lo sviluppo di un numero maggiore di casi d'uso e applicazioni.

  - La mancanza di uno schema predefinito consente di raccogliere e salvare i dati esattamente come sono, in modo semplice e veloce.
  - Nonostante sia impegnativo da gestire, un volume maggiore di dati consente migliori deduzioni e opportunità di analisi.

- **Svantaggi**:
  - L'ampio insieme di formati rende impegnativi l'analisi e l'utilizzo dei dati.
  - I formati non definiti dei dati richiedono strumenti specializzati per la loro gestione.

Infine, nei dati semi-strutturati rientrano tutti quei dati che non contengono informazioni strutturate secondo schemi fissi che non ammettono alcun tipo di modifica, ma che comunque seguono una struttura meno rigida e più tollerante nei confronti di possibili variazioni, aggiunte o omissioni di elementi. Questi dati sono spesso rappresentati tramite chiavi o marcatori che separano e organizzano le informazioni in modo semantico e gerarchico. Esempi di dati semi-strutturati possono essere i documenti scritti in formato "JavaScript Object Notation" (JSON), "eXtensible Markup Language" (XML), o "HyperText Markup Language" (HTML).

I dati che fanno parte di questa categoria possono essere facilmente integrati da diverse risorse e scambiati tra diversi sistemi. Inoltre, si adattano molto più facilmente alle diverse esigenze rispetto ai dati strutturati. Per esempio, volendo raccogliere dei dati da diversi utenti durante la compilazione di un form online, questi consentono di collezionare informazioni differenti, per numero e tipologia, per utenti diversi, senza la necessità di aggiornare lo schema ogni volta che viene richiesto un nuovo campo. Di conseguenza, aggiungere o rimuovere delle informazioni non impatta sulla funzionalità o sulle dipendenze di una piattaforma che sfrutta questa categoria di dati.

Per concludere questa Sezione, è stata inserita la Tabella 2.1 che riassume e schematizza le caratteristiche principali per le due tipologie di dati strutturati e non strutturati presentate e descritte sopra.

2.1.3 Riservatezza, Integrità, Disponibilità

"La cybersecurity è l'arte di proteggere reti, dispositivi e dati da accessi non autorizzati o da usi criminali e la pratica di garantire la riservatezza, l'integrità e la disponibilità delle informazioni"[5]. Questa definizione di cybersecurity data dalla Cybersecurity and Infrastructure Security Agency (CISA)[1], evidenzia come il rispetto dell'insieme di questi tre principi costituisca uno dei fondamenti della sicurezza informatica.

[1] Agenzia federale del Dipartimento della Sicurezza Interna degli Stati Uniti che opera per migliorare la sicurezza cibernetica e la protezione delle infrastrutture critiche della nazione dai rischi informatici in settori come energia, trasporti, comunicazioni e altri sistemi essenziali.

Caratteristica	Dati strutturati	Dati non strutturati
Formato	predefinito	nativo
Organizzazione	semplice	complessa
Analisi	semplice	complessa
Esempi	tabelle, basi di dati, fogli di calcolo	testo, audio, video, immagini
Approccio di immagazzinamento e gestione	Schema-on-write	Schema-on-read
Stima quantità	90%	10%

Tabella 2.1. Caratteristiche principali dei dati strutturati e non strutturati

In inglese, riservatezza, integrità e disponibilità si traducono rispettivamente in Confidentiality, Integrity e Availability, definendo la triade nota con l'acronimo "CIA".

Garantire la riservatezza delle informazioni significa proteggerle da accessi non autorizzati. Assicurare l'integrità dei dati vuole dire mantenerne l'accuratezza e la completezza rispetto alla loro forma originale, mantenendo una protezione delle informazioni da modifiche o cancellazioni improprie. Infine, rispettare il principio di disponibilità significa garantire agli utenti di poter accedere alle informazioni desiderate in qualsiasi momento risulti necessario.

Se una o più delle tre componenti della triade viene compromessa, la sicurezza dei dati viene violata. Questo può accadere in seguito ad eventi accidentali o in seguito ad attacchi malevoli.

Violazione dei dati

Una violazione di dati o "data breach" è un incidente o un attacco informatico che causa una compromissione della sicurezza delle informazioni. Le implicazioni possono essere diverse, dal furto di dati o di identità, all'accesso non autorizzato ai dati o la diffusione di informazioni confidenziali.

Le principali tipologie di violazione dei dati possono ricadere nella perdita delle informazioni, il cosiddetto "data leakage", o nel danneggiamento delle stesse. La perdita di dati può essere causata da violazioni interne o esterne alle organizzazioni. Spesso la violazione dei dati è causata dall'esterno con intenti malevoli, ma può anche accadere per disattenzione, negligenza o semplice incompetenza da parte di attori interni all'organizzazione.

Tra le principali cause di violazioni dei dati intenzionali dall'esterno ci sono:

- Malware
- Ransomware
- Social Engineering
- Phishing

La cause e di violazioni dei dati dall'interno possono essere intenzionali o accidentali. Alcuni esempi del primo caso sono:

- Modifica o cancellazione non autorizzata di informazioni per sabotaggio
- Copia non autorizzata di dati per scopi illeciti

Tra le violazioni non intenzionali, invece, si possono riscontrare spesso i seguenti casi::

- Errori di configurazione
- Abuso di privilegi
- Pubblicazione accidentale di contenuti
- Perdita o furto di dispositivi personali

Le violazioni causate da attori interni sono più comuni di quanto si pensi e sono quasi sempre una conseguenza di errori umani. Un report di Verizon [34] rivela che il 34% delle minacce nel settore finanziario e assicurativo proviene da fonti interne. L'errore umano è una delle principali cause di queste minacce e il più comune consiste nell'invio di dati riservati al destinatario sbagliato. La rilevazione degli incidenti interni di perdita di dati è più difficile, perché le violazioni interne coinvolgono spesso utenti che hanno accesso legittimo alle strutture e ai dati [4].

Per evitare la perdita di dati volontaria o involontaria possono essere impiegate varie misure di sicurezza, tra cui sistemi di prevenzione e rilevazione delle perdite di dati, che si occupano di impedire e riconoscere l'esposizione di tali informazioni. Tuttavia, oltre agli strumenti tecnologici, è essenziale anche sensibilizzare gli utenti sulla sicurezza sul posto di lavoro.

Le violazioni di dati mettono a rischio le informazioni critiche per le aziende e le informazioni che identificano delle persone (PII)[2]. La perdita e la compromissione di dati possono rappresentare per le aziende un costo importante in termini di reputazione, di operatività e di denaro.

IBM pubblica annualmente un report che porta alla luce dati molto interessanti sul costo medio di una violazione di dati per le organizzazioni. Nell'ultima edizione [14], dallo

[2]Personally Identifiable Information

studio di 553 organizzazioni colpite da violazioni di dati, è stato rivelato che il costo medio di una violazione di dati ha raggiunto il livello più alto di sempre nel 2023. Si tratta di una media di 4.45 milioni di dollari, rappresentando un aumento del 15.3% rispetto ai dati ottenuti nel rapporto del 2020. Un dato molto importante è che l'82% delle violazioni rilevate hanno coinvolto dati memorizzati nel cloud. Inoltre, il settore sanitario risulta il più colpito, seguito subito dopo dal settore finanziario.

Questi dati esprimono l'importanza e l'urgenza di implementare misure di sicurezza sempre più efficaci contro le nuove minacce.

La perdita economica, oltre ad essere conseguenza diretta della violazione dei dati, può essere una conseguenza indiretta delle multe e sanzioni che diversi organi possono applicare in caso di non conformità con le leggi in materia di tutela dei dati personali. Queste leggi regolamentano anche il comportamento da seguire in caso di avvenuta violazione dei dati. Ad esempio, secondo il Regolamento Generale sulla Protezione dei dati [8], il titolare del trattamento dei dati che viene a conoscenza di una violazione che ricade in delle casistiche ben stabilite, deve "senza ingiustificato ritardo e, ove possibile, entro 72 ore" notificare la violazione al Garante per la protezione dei dati personali. Il Garante può prescrivere misure correttive nel caso sia rilevata una violazione delle disposizioni o applicare sanzioni pecuniarie [31].

2.2 Quadro normativo

La crescente generazione di informazioni digitali ha creato l'esigenza di regolamentare diversi aspetti del trattamento dei dati, con particolare attenzione verso quelli personali e sensibili. Per garantire il rispetto dei diritti fondamentali delle persone e la sicurezza delle informazioni digitali, sono state elaborate numerose normative a livello nazionale e internazionale, che stabiliscono principi, regole e sanzioni per coloro che trattano i dati.

Le norme e gli standard per il trattamento dei dati variano a seconda del contesto e della finalità del trattamento stesso.

2.2.1 Regolamento Generale per la Protezione dei Dati

Nell'Unione Europea, il 2018 ha segnato una svolta fondamentale con l'applicazione del Regolamento Generale per la Protezione dei Dati [8], noto con l'acronimo GDPR[3], che rappresenta tuttora uno dei più rilevanti e completi riferimenti normativi in materia di protezione dei dati. Tutte le organizzazioni che gestiscono dati di cittadini o residenti dell'Unione Europea sono tenute a osservare le regole e i doveri stabiliti dal regolamento.

In esso, sono introdotti alcuni concetti e principi che meritano di essere menzionati poiché rilevanti per questo lavoro:

[3]General Data Protection Regulation

Definizione di dato personale (art. 4, punto 1): "qualsiasi informazione riguardante una persona fisica identificata o identificabile ("interessato"); si considera identificabile la persona fisica che può essere identificata, direttamente o indirettamente, con particolare riferimento a un identificativo come il nome, un numero di identificazione, dati relativi all'ubicazione, un identificativo online o a uno o più elementi caratteristici della sua identità fisica, fisiologica, genetica, psichica, economica, culturale o sociale;"

Particolari categorie di dati personali (dati sensibili) (art. 9, punto 1): "[...] dati personali che rivelino l'origine razziale o etnica, le opinioni politiche, le convinzioni religiose o filosofiche, o l'appartenenza sindacale, nonché trattare dati genetici, dati biometrici intesi a identificare in modo univoco una persona fisica, dati relativi alla salute o alla vita sessuale o all'orientamento sessuale della persona."

Definizione di dati relativi alla salute (art. 4, punto 15): "i dati personali attinenti alla salute fisica o mentale di una persona fisica, compresa la prestazione di servizi di assistenza sanitaria, che rivelano informazioni relative al suo stato di salute;"

Principio di limitazione della conservazione (art. 5, punto 1.e): "I dati personali sono conservati in una forma che consenta l'identificazione degli interessati per un arco di tempo non superiore al conseguimento delle finalità per le quali sono trattati; i dati personali possono essere conservati per periodi più lunghi a condizione che siano trattati esclusivamente a fini di archiviazione nel pubblico interesse, di ricerca scientifica o storica o a fini statistici, conformemente all'articolo 89, paragrafo 1, fatta salva l'attuazione di misure tecniche e organizzative adeguate richieste dal presente regolamento a tutela dei diritti e delle libertà dell'interessato («limitazione della conservazione»);"

2.2.2 Altre normative nazionali e internazionali sui dati

Il GDPR ha avuto una grande influenza anche nel contesto extra europeo, tanto che molte nazioni lo hanno usato come modello per formulare nuovi regolamenti locali o aggiornare quelli esistenti. É il caso, ad esempio, del Privacy Act 1988 australiano, per il quale sono state suggerite delle modifiche sulla base della regolamentazione europea e raccolte in un rapporto [7] che propone una revisione dell'atto già vigente. Il governo australiano ha pubblicato la sua risposta al rapporto il 28 settembre 2023, accettando la maggior parte delle raccomandazioni suggerite.

Nel Regno Unito, il Data Protection Act (DPA) del 2018 ha incorporato il GDPR nel diritto nazionale del Regno Unito a seguito della Brexit, introducendo alcune deroghe e modifiche. Il livello di protezione dei dati offerta dal DPA 2018 è essenzialmente sovrapponibile a quello garantito dal GDPR, tanto che la Commissione Europea ha riconosciuto

che i dati personali dei cittadini europei possono essere trasferiti liberamente dall'Unione Europea al Regno Unito, senza bisogno di ulteriori garanzie o deroghe.

Anche il Brasile si è ispirato al modello europeo, introducendo nel 2021 una legge sulla protezione dei dati personali chiamata Lei Geral da Proteção de Dados (LGPD), che presenta numerose analogie con il GDPR.

Altri esempi di regolamenti per la protezione dei dati sono:

Health Insurance Portability and Accountability Act (HIPAA)
che regola la protezione e la riservatezza dei dati sanitari personali negli Stati Uniti.

Personal Information Protection and Electronic Documents Act (PIPEDA)
che si applica in Canada alle organizzazioni del settore privato che raccolgono, utilizzano o divulgano informazioni personali nell'ambito di un'attività commerciale [30].

California Consumer Privacy Act (CCPA)
che concede ai consumatori maggior controllo sulle informazioni personali che le aziende raccolgono su di loro e e forniscono indicazioni su come applicare la legge [28].

La presenza di normative che cambiano da nazione a nazione, complica la posizione delle grandi aziende che operano a livello internazionale nella sfida di garantire la conformità. Il trasferimento dei dati da un Paese all'altro è infatti soggetto a condizioni di adeguatezza, soprattutto per i Paesi dove le norme sono più stringenti, come nel caso dell'Unione Europea. Per venire in contro a questi limiti sono state elaborate delle soluzioni e degli accordi, come il Data Privacy Framework (DPF) UE-USA, che riguarda il trasferimento dei dati personali dall'Unione Europea agli Stati Uniti. Il DPF prevede che le organizzazioni statunitensi che trattano dati personali provenienti dall'UE si impegnino a rispettare una serie di principi e requisiti in materia di privacy equivalenti a quelli stabiliti dal GDPR.

Oltre alle normative menzionate, esistono diversi standard di sicurezza globali che stabiliscono delle linee guida per la protezione e la conservazione di alcuni dati particolari.

Il Payment Card Industry Data Security Standard (PCI-DSS) è uno standard di conformità che stabilisce i requisiti minimi per la protezione dei pagamenti dei consumatori e dei dati finanziari. Il PCI-DSS si applica a tutte le entità che memorizzano, elaborano o trasmettono dati di pagamento. Il suo scopo è di ridurre il rischio di frodi, furti e perdite di dati sensibili, garantendo che le entità che trattano i dati di pagamento adottino misure adeguate per prevenire, rilevare e rispondere agli incidenti di sicurezza. I requisiti che queste entità devono garantire riguardano aspetti come la protezione dei dati memorizzati, la crittografia delle trasmissioni, la gestione delle vulnerabilità, l'implementazione di controlli di accesso, il monitoraggio e il test delle reti e la definizione di politiche di sicurezza.

Un altro esempio è la norma ISO/IEC 27018:2020 [9], un codice di condotta per la sicurezza delle informazioni con indicazioni specifiche per i controlli e le misure di protezione da applicare nell'ambiente del cloud computing pubblico in cui si elaborano le Informazioni di Identificazione Personale (PII). È un'estensione della ISO 27001 e della ISO 27002 e fornisce controlli di sicurezza aggiuntivi.

Legislazione italiana

In Italia la gestione dei dati è regolamentata da diverse normative, tra cui il decreto legislativo del 30 giugno 2003, n.196 "Codice in materia di protezione dei dati personali", detto anche Codice Privacy, che è stato integrato negli anni per l'adeguamento al GDPR.

In alcuni casi, le norme diventano più specifiche soprattutto per quanto riguarda la raccolta e la conservazione delle informazioni.

Ad esempio, dal 1 gennaio 2022, per tutte le aziende e gli Enti Pubblici è obbligatorio redigere e conservare il proprio Manuale di Conservazione, anche se il servizio di conservazione è affidato a un fornitore esterno. Questo è quanto previsto dalle Linee Guida [32] dell'Agenzia per l'Italia Digitale (AgID) sulla formazione, gestione e conservazione dei documenti informatici. Il Manuale regola la vita del documento informatico, dalla sua creazione alla sua conservazione e accessibilità mediante sistemi informatici.

Il decreto legislativo del 21 novembre 2007 n.231 afferma che i soggetti obbligati devono conservare diverse tipologie di dati indicate nel decreto per un periodo di 10 anni per "assolvere gli obblighi di adeguata verifica della clientela affinché possano essere utilizzati per qualsiasi indagine su eventuali operazioni di riciclaggio o di finanziamento del terrorismo o per corrispondenti analisi effettuate dalla UIF o da qualsiasi altra Autorità competente" [1]. Questo periodo può essere prorogato nel caso in cui venisse richiesto da autorità competenti.

2.3 Rischio informatico

Il concetto di rischio informatico, noto comunemente come "Cyber Risk" in inglese, costituisce uno degli elementi chiave su cui si basa il lavoro svolto in questa tesi.

Svariate organizzazioni, che si occupano di sviluppare e promuovere degli standard in ambito tecnologico, di sicurezza informatica e di gestione dei rischi connessi, hanno formulato la loro definizione di rischio informatico. Qui di seguito, sono menzionate due di queste definizioni, fornite dal NIST[4] e dall'ISO[5]:

Definizione di Rischio secondo il NIST :

"Una misura del grado con cui un'entità è minacciata da una potenziale circostanza

[4]National Institute of Standards and Technology

[5]International Organization for Standardization

o evento e una combinazione di: (i) impatti avversi che si verrebbero a creare se la circostanza o l'evento si verificasse; e (ii) la probabilità di occorrenza" [29].

Definizione di Rischio secondo l'ISO 31000 :

"Effetto di incertezza sugli obiettivi, dove per effetto si intende una deviazione dal previsto che può essere positivo, negativo o entrambi, e può indirizzare, creare o determinare opportunità e minacce. Gli obiettivi possono avere diversi aspetti e categorie e possono essere applicati a diversi livelli. Il rischio si esprime in funzione di fonti di rischio, eventi, conseguenze e probabilità" [16].

In altri termini, il rischio informatico è la probabilità di una organizzazione di essere esposta o subire dei danni (perdite economiche, di dati, di reputazione) a causa di un attacco informatico o di una violazione dei dati.

Oggi, quasi tutte le aziende si affidano in buona parte, se non interamente, a sistemi informatizzati, comportando un'elevata esposizione agli attacchi più comuni che possono impattare in modo molto negativo se non vengono prese le dovute precauzioni.

Tra le precauzioni necessarie per limitare gli attacchi o mitigarne gli effetti, rientrano sicuramente tutte quelle tecniche di protezione che includono l'uso di firewall, antivirus, autenticazione a due fattori, crittografia dei dati, backup regolari dei dati e formazione degli utenti su come riconoscere e prevenire potenziali minacce come phishing e malware. È importante notare che nessuna singola tecnica può garantire una protezione completa, quindi l'uso combinato di diverse tecniche di protezione può fornire una difesa più robusta.

Applicare tutte le tecniche di sicurezza necessarie a ogni possibile vettore di attacco informatico potrebbe ridurre significativamente il rischio di subire attacchi. Nella pratica, però, l'implementazione di tutti i sistemi di sicurezza potrebbe rivelarsi impraticabile a causa dei costi associati.

Infatti, in assenza di una guida o di un'analisi, le organizzazioni potrebbero concentrare i propri sforzi economici in modo inefficiente. Pertanto, è fondamentale stabilire una strategia di gestione del rischio e condurre un'analisi per identificare e dare la giusta priorità alle aree che necessitano di maggior protezione.

Una gestione efficiente del rischio, per essere definita tale, deve comprendere tutta una serie di azioni che devono essere svolte in modo coordinato e consistente. Alcune delle azioni chiave sono riportare di seguito:

- Sviluppo di politiche e di strumenti affidabili per valutare il rischio.
- Identificazione di nuovi rischi o di nuove regolamentazioni.
- Identificazione di debolezze interne all'organizzazione, come la mancanza di autenticazione a due fattori.
- Mitigazione dei rischi informatici, eventualmente attraverso programmi di formazione o nuove politiche e controlli interni.

- Verifica dei sistemi di sicurezza nel complesso.
- Stesura della documentazione sulla gestione del rischio e della sicurezza

Tra le azioni elencate sopra, quelle di maggiore impatto che vanno a delineare il profilo di rischio dell'organizzazione durante il processo di gestione, vengono raggruppate sotto il nome di "Risk Assessment". Questo insieme di azioni verte ad identificare, esaminare e valutare tutte le possibili minacce e vulnerabilità che possono influire sulla capacità di un' organizzazione di svolgere le sue attività.

Il "Risk Assessment" è composto da due parti principali: l'identificazione del rischio e l'analisi del rischio. In particolare, durante questo processo devono essere considerati ed identificati vari tipi di pericoli e rischi (non solo quelli legati alla sicurezza delle informazioni e delle infrastrutture informatiche) e va compilato un registro dei rischi che aiuta a documentare e categorizzare i risultati del processo di identificazione. Successivamente, bisogna considerare in che modo questi rischi possono danneggiare l'organizzazione e quali sono i potenziali obiettivi al fine di definire gli eventuali esiti possibili. Infine, si esamina ogni rischio identificato e gli si assegna un punteggio utilizzando uno o entrambi i due sistemi di punteggio: quantitativo o qualitativo. Questi punteggi aiutano a definire le priorità dei rischi, in modo da sapere quali affrontare per primi e quali sono i modi migliori per farlo.

Nello studio sul rischio condotto in questo lavoro, è stata svolta un'analisi qualitativa. Di seguito vengono elencate le principali differenze tra le due categorie di analisi del rischio:

Analisi del rischio qualitativa :
questo metodo di analisi categorizza i rischi sulla base di una combinazione della loro probabilità di manifestarsi e dell'impatto potenziale che potrebbero avere. La classificazione si basa su valutazioni soggettive e, di conseguenza, può essere influenzata dalla percezione e dall'esperienza di chi la conduce. A ogni rischio viene assegnato un grado su una scala che varia tipicamente da "molto alto" a "molto basso". Questa metodologia è semplice e rapida da implementare, rendendola adattabile a qualsiasi organizzazione e applicabile a qualsiasi tipo di risorsa che si intende valutare.

Analisi del rischio quantitativa :
Questo metodo di analisi si avvale di tecniche matematiche e statistiche per calcolare l'impatto finanziario dei rischi. Fornisce risultati oggettivi derivanti da calcoli, rendendolo più preciso e dettagliato rispetto al precedente. Tuttavia, la sua complessità non lo rende adatto per essere applicato a un gran numero di risorse da valutare. Poiché si concentra sugli impatti economici, questa analisi risulta particolarmente appropriata per facilitare la comprensione dei rischi a livelli manageriali e di coloro che non possiedono competenze tecniche.

Sia l'analisi del rischio quantitativa che quella qualitativa presentano vantaggi e limitazioni, e dovrebbero essere impiegate in maniera appropriata. In generale, può essere

efficace condurre inizialmente un'analisi del rischio qualitativa e utilizzare una matrice dei rischi per facilitare l'identificazione delle minacce a priorità più alta. Successivamente, si può procedere con l'analisi del rischio quantitativa per le minacce più rischiose.

È fondamentale ricordare che i sistemi, in particolare quelli tecnologici, sono soggetti a un'evoluzione costante ed è essenziale che le valutazioni dei rischi siano sottoposte a revisioni periodiche.

2.4 Cloud Computing

Il cloud computing consiste nella distribuzione on-demand tramite Internet di risorse IT [3] come ad esempio: applicazioni, server fisici e virtuali, archiviazione dati, strumenti di sviluppo, funzionalità di rete [13].

I servizi cloud sono gestiti da fornitori di terze parti detti Cloud Service Providers (CSP), che offrono l'accesso a pagamento alle risorse ospitate in data center remoti. Per questo motivo, la scelta delle risorse a cui accedere può essere effettuata in base alle proprie necessità, in qualsiasi momento e da qualsiasi luogo. Inoltre, i CSP offrono agli utenti la possibilità di personalizzare l'offerta scelta, proponendo diversi modelli di pagamento dei servizi in base alle proprie esigenze e all'utilizzo effettivo delle risorse. Le opzioni spaziano da pagamenti anticipati a cadenza regolare, come nel caso degli abbonamenti, a pagamenti basati sul consumo effettivo, come nella formula "Pay As You Go" (PAYG)[6].

La tecnologia cloud consiste, dunque, in una sorta di "noleggio" di risorse IT e infrastrutture condivise, contrapponendosi al modello on-premise[7]. Optando per una soluzione basata sul cloud, le aziende possono ridurre i tempi e i costi richiesti per la gestione delle soluzioni locali, evitando l'investimento nell'acquisto di infrastrutture necessarie e nella loro configurazione e manutenzione.

Ricapitolando, i motivi per cui l'adozione di soluzioni basate sul cloud può rivelarsi conveniente per le organizzazioni sono molteplici:

- Semplificazione della gestione delle risorse
- Controllo dei costi
- Scalabilità su richiesta in base al carico di lavoro
- Flessibilità

[6] "Paghi quello che usi", anche definito come "pay as you use" o "pay per use". Si tratta di un modello tariffario che prevede il pagamento solo dei servizi che vengono utilizzati, per il tempo in cui vengono utilizzati.

[7] Approccio in cui tutte le risorse informatiche utilizzate da un'organizzazione sono "site in loco", ossia vengono implementate e gestite all'interno dell'organizzazione stessa. In questo caso, l'azienda è responsabile dell'acquisto, la manutenzione, l'aggiornamento e la gestione di tali risorse.

L'insieme dei vantaggi precedentemente elencati, ha influito sulla decisione di molte organizzazioni di migrare da infrastrutture on-premise ad altre on-cloud. Tuttavia, è essenziale che le aziende tengano in considerazione che nel mondo del cloud computing ci sono ancora diverse sfide aperte legate alla sicurezza, in modo da poter effettuare una scelta consapevole e fondata sulle proprie esigenze.

In Tabella 2.2 è riassunto un confronto relativo ad alcuni degli aspetti principali da valutare nella scelta tra una soluzione on-cloud e una on-premise. In sintesi, entrambe le

	ON-PREMISE	ON-CLOUD
AUTONOMIA	Controllo totale sui propri dati e sui sistemi. Il data center è l'architettura IT sono totalmente di proprietà dell'organizzazione.	Server remoti gestiti da un fornitore esterno.
COSTI	Alti costi iniziali per l'acquisto di un'infrastruttura propria. Costi relativi alla gestione delle risorse: configurazione, manutenzione.	Spese distribuite nel tempo e proporzionate all'utilizzo effettivo delle risorse.
SCALABILITÀ	Disponibilità di risorse limitata che tipicamente non può far fronte a carichi di lavoro imprevisti.	Disponibilità elevata di funzionalità e risorse per soddisfare nuovi e più pesanti carichi di lavoro.
SICUREZZA	Pieno controllo dei dati che non vengono ceduti a terzi, in particolare i dati sensibili. Rischio di soluzioni di sicurezza obsolete.	È possibile usufruire di una tecnologia in costante aggiornamento e per questo potenzialmente più sicura. Sicurezza affidata a un team esterno, per cui si perde il controllo totale dei dati.

Tabella 2.2. Confronto degli aspetti principali delle soluzioni on-premise e on-cloud

soluzioni hanno i loro vantaggi e svantaggi. Non c'è una soluzione migliore in assoluto tra le due proposte, la scelta di quale adottare dipende dalle esigenze specifiche dell'azienda e non si può escludere la possibilità di mantenere entrambe le soluzioni da impiegare per scopi differenti tra loro.

2.4.1 Modelli di cloud computing

Nel cloud computing, un provider esterno si occupa della gestione di un servizio per conto dell'utente. Questo permette all'utente di focalizzare le proprie risorse su attività di maggiore interesse, affidando la gestione di parte o di tutti i componenti dell'infrastruttura. A seconda del livello di gestione che l'utente decide di delegare, esistono diversi tipi di servizi cloud "as-a-Service".

I tre modelli principali di cloud computing, che si distinguono per il modo in cui i servizi vengono forniti, sono:

Infrastructure as a Service (IaaS): questo modello elimina la necessità per l'utente di gestire, mantenere o aggiornare la propria infrastruttura, delegando questo compito ad un fornitore terzo. Il provider offre archiviazione, networking e virtualizzazione su richiesta, mentre l'utente rimane responsabile per il sistema operativo, il middleware, le macchine virtuali e qualsiasi applicazione o dato correlato.

Platform as a Service (PaaS): con questo modello è possibile creare, sviluppare e distribuire le proprie applicazioni avendo a disposizione piattaforme e ambienti cloud preconfigurati e pronti all'uso. L'utente non deve preoccuparsi dell'infrastruttura sottostante, la cui gestione e manutenzione è sotto la responsabilità del provider del servizio.

Software as a Service (SaaS): è un modello che fornisce un'applicazione completa utilizzabile dagli utenti. I prodotti sono gestiti integralmente dal fornitore di servizi, inclusi tutti gli aggiornamenti, le correzioni di bug e la manutenzione generale. Gli utenti non hanno bisogno di scaricare o installare software sui propri dispositivi, ma possono accedervi direttamente dal Web.

I modelli di servizio IaaS, PaaS e SaaS non si escludono tra loro ed è molto comune per le aziende scegliere di adottare più di uno di questi modelli di servizio.

2.4.2 Strumenti di archiviazione e collaborazione su Cloud

Il cloud computing si è affermato come uno strumento essenziale per le aziende, soprattutto per quanto riguarda l'archiviazione e la condivisione di dati. Le piattaforme di collaborazione permettono agli utenti di lavorare simultaneamente sui file, apportando modifiche in tempo reale e assicurando che ogni membro del team abbia accesso alla versione più recente del progetto. La possibilità di avere un spazio di archiviazione centralizzato facilita l'accesso ai documenti da qualsiasi luogo e in qualsiasi momento, riducendo la necessità di allegati via email.

Gli strumenti di collaborazione basati sul cloud non si limitano alla condivisione di documenti, ma facilitano anche l'interazione tra i membri del team attraverso servizi di messaggistica istantanea. La possibilità di avere calendari condivisi consente, inoltre, di pianificare riunioni ed eventi.

In generale, queste soluzioni sono progettate per migliorare la cooperazione tra individui che lavorano insieme a vari progetti.

Alcune delle piattaforme collaborative basate sul cloud più popolari includono Cisco Webex, Microsoft Teams, Meta Workplace, Google Workspace, Dropbox e Slack.

Capitolo 3

Obiettivo

In un contesto aziendale in cui vengono trattati dati sensibili, la protezione di tali informazioni rappresenta una priorità strategica. Questo diventa ancor più rilevante quando i dati vengono salvati su piattaforme cloud, che ne favoriscono la condivisione e la diffusione.

La tematica centrale di questo lavoro di tesi è stata la protezione dei dati aziendali ospitati in archivi cloud durante specifiche fasi del loro ciclo di vita. In particolare, ci si è concentrati sull'implementazione di controlli di sicurezza sulla piattaforma cloud SharePoint Online, un software di collaborazione e incluso nei servizi offerti da Microsoft 365. Nonostante i controlli implementati siano stati progettati specificamente per lo strumento menzionato, l'approccio utilizzato per la loro realizzazione può essere adattato ed esteso per essere applicato anche ad altri strumenti di collaborazione sul cloud.

SharePoint Online è uno strumento che le organizzazioni possono utilizzare per creare i cosidetti Siti SharePoint. Questi possono essere siti Web tradizionali o spazi dedicati all'archiviazione e alla condivisione di informazioni, accessibili in qualsiasi momento. Infatti, Microsoft promuove SharePoint Online come uno strumento per permettere la condivisione di contenuti e collaborare "senza barriere in tutta l'organizzazione" [26]. Sebbene questo aspetto possa essere utile per supportare il lavoro in team, può rappresentare un rischio se lo strumento non viene utilizzato correttamente o se non vengono applicate le giuste restrizioni alla condivisione di dati riservati.

Gli strumenti di collaborazione vengono utilizzati all'interno delle organizzazioni da un gran numero di dipendenti con diversi livelli di competenze e consapevolezza riguardo la sicurezza informatica. Oltre a ciò, esistono normative specifiche sul trattamento dei dati, introdotte nel capitolo 2.2, di cui i dipendenti devono essere a conoscenza in base al tipo di dati che trattano. Sebbene sia essenziale fornire una formazione adeguata del personale all'interno di un'organizzazione, ciò potrebbe non essere sufficiente a garantire un livello di protezione adeguato e costante. Infatti, una gestione manuale che lascia tutte le responsabilità all'utente può facilmente essere soggetta ad errori umani.

Il lavoro descritto in questa tesi è stato svolto con l'obiettivo di affrontare le sfide legate

a una protezione efficace ed efficiente dei dati. Questo viene fatto attraverso l'implementazione di soluzioni automatiche che ottimizzano i controlli di sicurezza, supportano l'utente e colmano le eventuali lacune causate da errori umani, dimenticanze o una gestione inadeguata degli strumenti aziendali.

È importante sottolineare che, nell'azienda dove è stato condotto questo lavoro, non erano presenti tali controlli automatici. Pertanto, l'introduzione di queste soluzioni ha rappresentato un passo significativo verso il potenziamento della sicurezza dei dati dell'azienda. Inoltre, le strategie utilizzate in questo contesto possono risultare vantaggiose in particolare per le aziende che gestiscono grandi quantità di dati di varia natura e che, sono soggette al rispetto di specifici criteri di sicurezza informatica e vincoli normativi.

Per raggiungere lo scopo di potenziare la sicurezza dei dati aziendali, durante specifiche fasi del loro ciclo di vita nel cloud, sono stati intrapresi una serie di interventi, sfruttando i vantaggi degli automatismi. Grazie al lavoro svolto, è stato possibile raggruppare gli

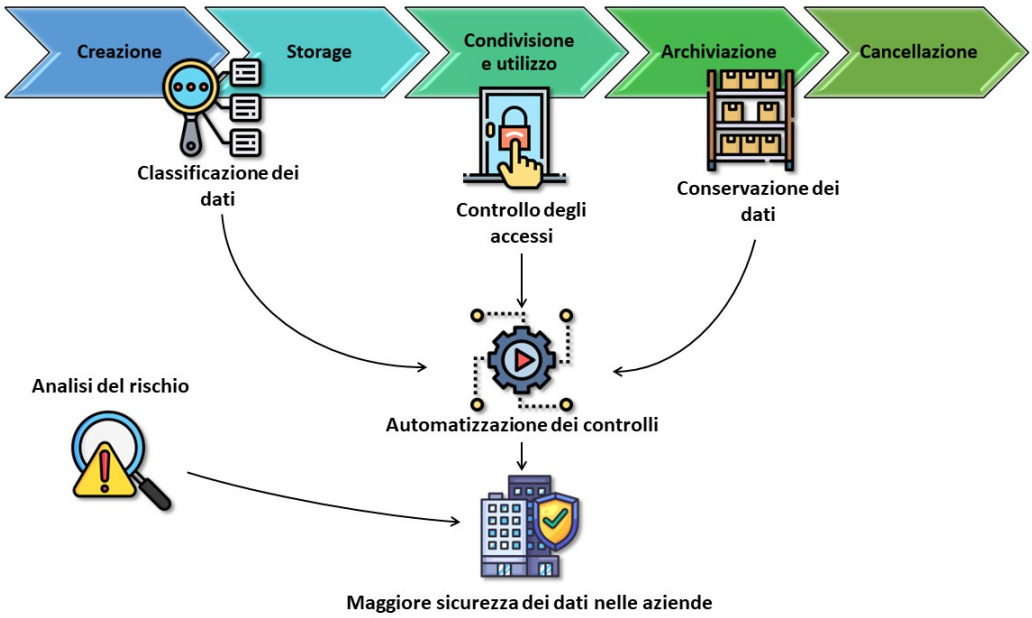

Figura 3.1. Rappresentazione delle azioni svolte per il raggiungimento dell'obiettivo

interventi per aree di competenza, come mostrato in Figura 3.1, analizzando per ciascuna le peculiarità introdotte:

Classificazione dei dati in base al livello di riservatezza del contenuto.
Questa operazione consiste nell'assegnare una categoria a ciascuno dei file ospitati sui siti SharePoint dell'organizzazione in base al livello di riservatezza del contenuto. In altre parole, si tratta di identificare informazioni riservate, come possono essere

quelle personali e sensibili, e applicare un marcatore sui singoli file in base a queste caratteristiche.

L'adozione di un sistema automatico per la classificazione delle informazioni permette di:

- Fornire agli utenti indicazioni sulla natura del contenuto dei file, permettendo loro di gestirli e proteggerli in modo adeguato.
- Categorizzare in poco tempo grandi quantità di file.
- Classificare i file che non sono stati categorizzati manualmente dagli utenti, sia per dimenticanza sia perché sono stati ottenuti da fonti esterne all'organizzazione.
- Assegnare ai file dei marcatori che i sistemi aziendali di DLP già in uso possono riconoscere, al fine di prevenire la fuga di informazioni riservate.

Controllo dei permessi di accesso

L'attività consiste nel ricavare informazioni relative alla concessione dei permessi di accesso ai siti SharePoint aziendali. In particolare, il monitoraggio si concentra sulla ricerca di situazioni di potenziale rischio derivanti da una cattiva gestione dello strumento di collaborazione. Ad esempio, si controlla il numero di utenti dotati di permessi che consentono il controllo totale del sito e del suo contenuto, intervenendo nel caso in cui tale numero superi un determinato limite. La gestione automatica di questo monitoraggio consente di raggiungere i seguenti obiettivi:

- Aumentare l'efficienza nei controlli. Il sistema automatico, infatti, permette di individuare rapidamente le situazioni di potenziale rischio.
- Riprodurre il controllo con facilità e in qualsiasi momento.
- Assicurare la conformità alle normative sulla privacy monitorando le situazioni in cui l'accesso ai dati avviene in maniera incontrollata o da un numero eccessivo di persone senza una giustificazione valida.
- Eseguire controlli programmati con regolarità, garantendo che le informazioni sui permessi di accesso siano sempre aggiornate.

Sviluppo di una strategia di conservazione dei dati

Questa attività riguarda la progettazione di una strategia per determinare e applicare un periodo minimo di conservazione ai dati, in linea con i requisiti normativi e la natura dell'informazione trattata. Tale politica è indispensabile per evitare di incorrere in sanzioni legali e per assicurare che le informazioni siano disponibili per tutto il tempo in cui potrebbe essere necessario accedervi. Grazie all'integrazione degli automatismi è possibile ottenere i seguenti obiettivi:

- Prevenire la cancellazione accidentale di file che devono essere conservati.

- Migliorare la conformità alle normative vigenti.
- Stabilire le azioni da intraprendere automaticamente alla fine del periodo di conservazione. In questo modo si può prevenire l'eventualità di dimenticare l'esistenza di file che potrebbero essere eliminati, assicurando una gestione dello spazio di archiviazione più efficace.

A completamento del lavoro e per garantire una più valida valutazione dell'operato, è stata condotta una analisi di rischio per definire l'impatto di eventi avversi ante intervento e dopo lo studio.

L'analisi preliminare può essere soggetta a modifiche in base al numero di vulnerabilità che, dopo un'attenta valutazione, vengono effettivamente ritenute pericolose. Di conseguenza, in questo lavoro viene presentata anche una stima del rischio che ci si aspetta di avere a seguito dell'implementazione dei controlli proposti. Questo processo aiuta a valutare l'efficacia delle soluzioni fornite.

Capitolo 4

Architettura del Sistema

Nel presente capitolo, viene fornita una descrizione del flusso di lavoro e una presentazione degli strumenti utilizzati. Questo contenuto serve da guida per la comprensione della struttura del lavoro svolto, offrendo una visione chiara dei passi necessari per il conseguimento degli scopi prefissati.

4.1 Modellazione del flusso di lavoro

Nella fase iniziale del progetto di tesi, si è scelto di pianificare il lavoro da svolgere attraverso la definizione di un flusso di attività, al fine di avere una traccia chiara delle azioni da compiere. I processi sono stati delineati utilizzando delle rappresentazioni in BPMN, acronimo di "Business Process Model and Notation", che in italiano si traduce in "Notazione e Modellazione dei Processi Aziendali".

La notazione BPMN, definita dall'Object Management Group (OMG)[1] ed attualmente disponibile nella versione 2.0.2 [12], costituisce uno standard ampiamente utilizzato dalle aziende di tutto il mondo per la documentazione e l'ottimizzazione dei processi aziendali sottoforma di rappresentazione grafica.

Il principale vantaggio che questo metodo offre è la possibilità di generare diagrammi di flusso personalizzati sulle esigenze specifiche delle organizzazioni che siano intuitivi per tutti gli interessati all'interno dell'azienda, indipendentemente dal loro grado di competenza tecnica. Infatti, la rappresentazione visuale favorisce la comprensione dei processi, mentre l'utilizzo di una notazione universalmente accettata agevola la comunicazione tra diversi team di una stessa organizzazione e tra organizzazioni differenti.

I diagrammi in notazione BPMN si compongono di una serie di elementi e simboli per delineare il flusso di azioni da eseguire in modo sequenziale. La notazione definisce una

[1] Consorzio internazionale per gli standard tecnologici ad adesione libera e senza fini di lucro.

vasta gamma di simboli standardizzati e la scelta di quelli da impiegare dipende dalle esigenze specifiche del processo che si sta modellando.

In Figura 4.1 vengono riportati solo alcuni tra i principali elementi base e relativi simboli che è possibile utilizzare per la costruzione di un diagramma in BPMN: eventi, attività, gateways, flussi di sequenza e di messaggi, pool. In particolare, questi simboli

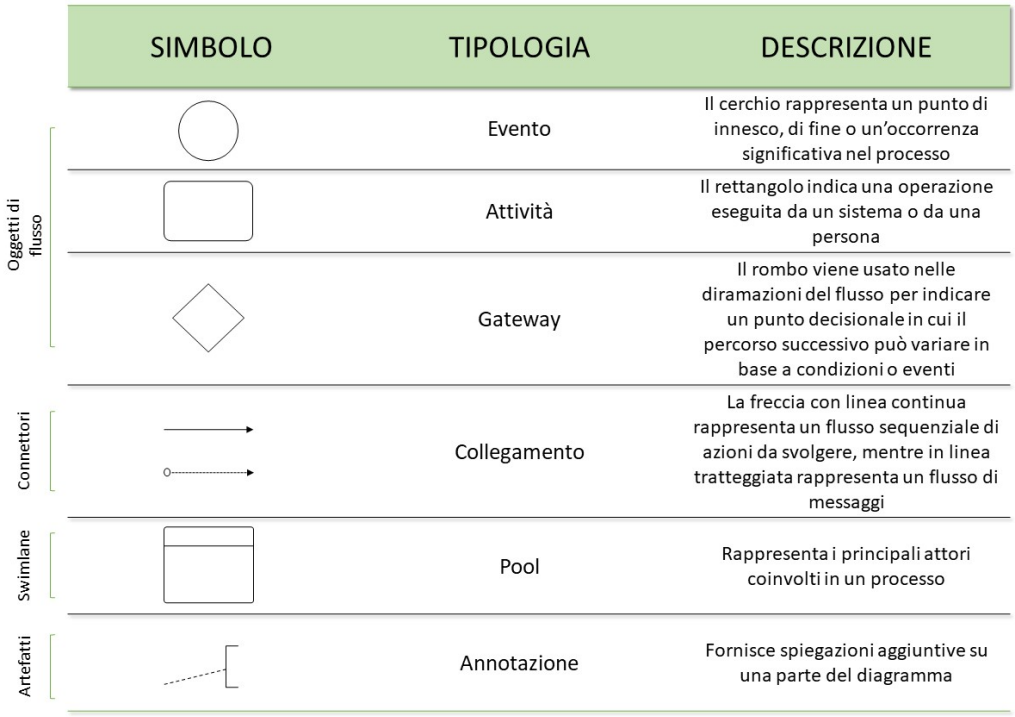

Figura 4.1. Descrizione dei simboli principali di un diagramma BPMN

sono quelli che sono stati impiegati per la realizzazione dei processi a cui si sta facendo riferimento. Come si può notare in Figura 4.1, i primi tre simboli appartengono alla categoria di elementi base chiamata "Oggetti di flusso", le frecce fanno parte dei "Connettori", mentre gli ultimi due simboli appartengono rispettivamente agli elementi base chiamati "Swimlane" e "Artefatti".

I simboli descritti in precedenza si suddividono in diverse sottoclassi, assumendo un significato più specifico in base ad altri simboli che possono essere contenuti al loro interno. La Figura 4.2 mostra un elenco di alcune di queste sottoclassi con i relativi simboli impiegati per la modellazione dei processi che saranno descritti nelle sezioni successive di questo capitolo.

Per la realizzazione di diagrammi in BPMN sono disponibili diversi programmi, molti dei quali utilizzabili online gratuitamente con una serie di funzionalità ridotte, ma che

si rivelano sufficienti per la creazione di diagrammi non troppo complessi. Alcuni tra

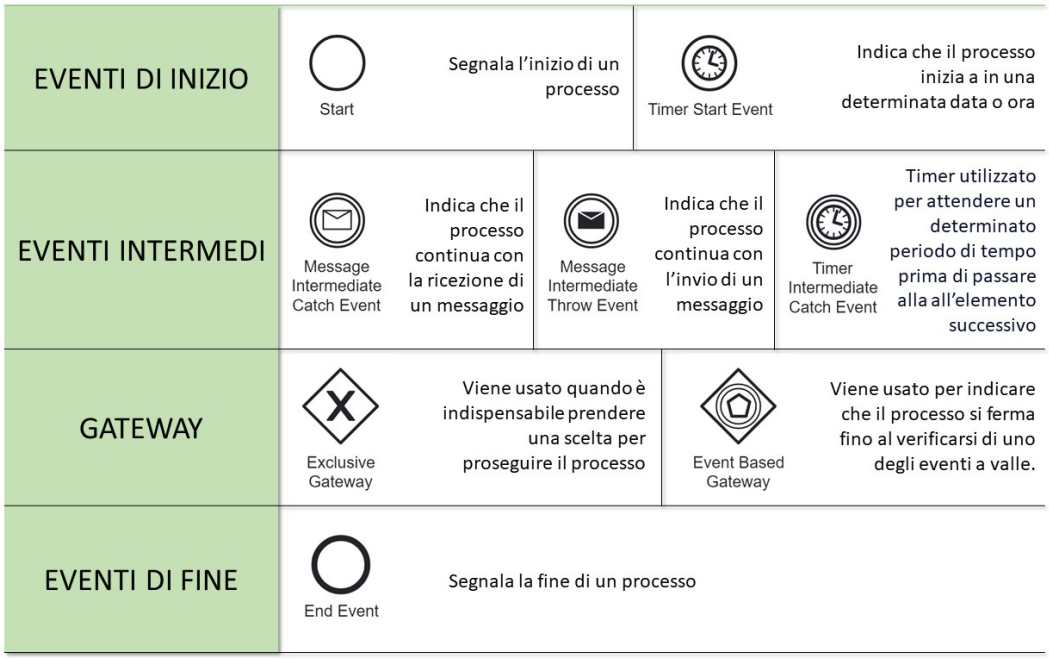

Figura 4.2. Sottoclassi di simboli in BPMN utilizzate nella modellazione dei processi

i più noti siti che forniscono questo tipo di strumenti sono: Lucidchart[2], Camunda[3] e BPMN.io[4]. Tra questi, Camunda è lo strumento che è stato utilizzato per la modellazione dei processi presentati nelle sezioni successive di questo capitolo.

4.1.1 Processo di classificazione

Come anticipato nei precedenti capitoli, la classificazione dei file è essenziale per definire le più adeguate misure di sicurezza e di gestione dei file. Per questa ragione, assume particolare importanza il processo di classificazione dei dati non strutturati tramite etichette di riservatezza.

Un'etichetta è un insieme di metadati associati a un file che consente di definirne il grado di riservatezza, la categoria di appartenenza o altre informazioni significative che lo

[2] https://www.lucidchart.com/

[3] https://camunda.com/bpmn/tool/

[4] https://bpmn.io/

caratterizzano. Questi dati categorizzano e descrivono il file in accordo a specifici criteri stabiliti dall'utente o dall'organizzazione. L'utilizzo delle etichette, dunque, permette una gestione più efficace dei file, agevolando il raggruppamento di documenti affini per tipologia di dati o livello di sicurezza da applicare e garantendo una protezione adeguata quando le etichette vengono applicate per limitare l'accesso a contenuti riservati.

Esistono diversi programmi che offrono agli utenti la possibilità di etichettare i file presenti sugli spazi di archiviazione e di condivisione in cloud. Ad esempio, in Google Drive è possibile identificare informazioni sensibili, assegnare loro delle etichette di riservatezza e, in base a queste, creare un criterio per la prevenzione della perdita di dati (DLP)[5], come viene descritto nella guida ufficiale [11]. Anche Dropbox offre una funzionalità di classificazione dei dati che permette di assegnare automaticamente ai file delle etichette per evidenziare la presenza di dati personali al loro interno.

Tuttavia, data l'ampia diffusione dei servizi di Microsoft 365 nei contesti aziendali, in questo progetto di tesi è stato utilizzato un prodotto offerto nella suite di servizi Microsoft, chiamato Microsoft Purview, un servizio di data governance che offre diverse funzionalità, molte delle quali discusse da Ahmad et al. [2]. Tra queste, quelle fornite dal SaaS Microsoft Information Protection (MIP) [21] sono state rilevanti per lo scopo della tesi, in particolare il riconoscimento di informazioni sensibili all'interno di file testuali e la possibilità di contrassegnare e proteggere i documenti mediante l'applicazione delle cosiddette "etichette di riservatezza" [20].

Informazioni più approfondite sull'utilizzo di questo strumento saranno esaminate in dettaglio nella Sezione 5.1.

Il diagramma, in Figura 4.3, illustra il processo di creazione delle policy[6] di etichettatura automatica per la classificazione dei dati.

Come si evince dalla rappresentazione, è necessario dapprima associare i tipi di informazioni riservate (SIT)[7], predefinite e messe a disposizione dallo strumento [25], alle etichette di riservatezza disponibili esistenti nell'organizzazione. I tipi di informazioni riservate sono classificatori basati su pattern [24] in grado di rilevare informazioni sensibili all'interno di documenti di tipo testuale. I SIT comprendono una lista di informazioni standard già pronte all'uso ma, in base alle necessità, possono essere creati su richiesta dall'amministratore di sistema nuovi elementi.

Durante questa fase, è importante stabilire una corrispondenza precisa tra le politiche definite in termini di classificazione dei dati e le varie tipologie di dati effettivamente trattate. Questo permette di associare a ciascun tipo di dato, in base al grado di riservatezza, l'etichetta più adatta tra quelle già esistenti nell'organizzazione.

[5] Data Loss Prevention

[6] regole o procedure che vengono applicate per classificare automaticamente i dati in base a determinati criteri o condizioni

[7] Sensitive Information Type

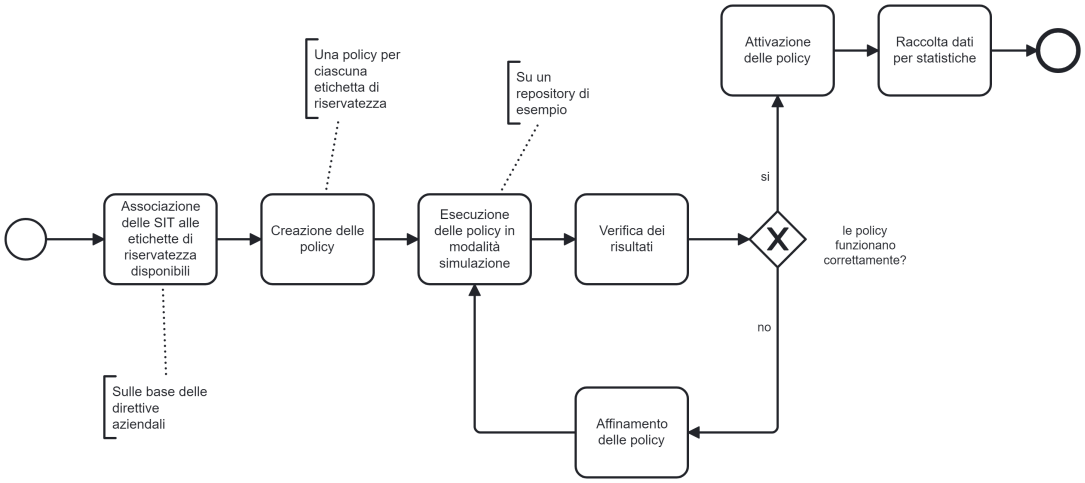

Figura 4.3. Diagramma in BPMN del processo di classificazione

Una volta definita tale associazione, si procede con la creazione di una policy per ciascuna etichetta di riservatezza individuata. L'efficienza delle policy può essere testata attraverso la "modalità simulazione" offerta dallo strumento. Questa funzionalità consente di riprodurre il funzionamento della policy e valutarne i risultati, prima di attivarla definitivamente. L'utilizzo della simulazione è particolarmente vantaggioso per esaminare i file che incontrano i criteri definiti nella policy, identificare potenziali falsi positivi e perfezionare la costruzione dei criteri per soddisfare al meglio le esigenze specifiche.

Una volta raggiunti dei risultati soddisfacenti dalle simulazioni, si può procedere con l'attivazione effettiva delle policy sui siti SharePoint di interesse. Questa fase permette di applicare le policy di classificazione automatica ai file presenti sui siti SharePoint e di monitorare i cambiamenti delle etichette applicate. In questo modo, si possono raccogliere dati statistici dettagliati sui file che sono stati etichettati per la prima volta e su quelli che hanno subito una modifica dell'etichetta assegnata. Questi dati si possono rivelare particolarmente utili per analizzare il comportamento degli utenti e per valutare il livello di protezione dei dati aziendali.

4.1.2 Processo di controllo dei permessi di accesso

Il processo presentato in questo paragrafo delinea le attività da eseguire per implementare un controllo automatico sui permessi di accesso ai siti SharePoint Online.

La gestione delle autorizzazioni di SharePoint avviene generalmente attraverso un insieme di gruppi di autorizzazioni presenti all'interno di un sito che identificano dei ruoli. I ruoli principali in SharePoint sono "Proprietari" o "Owners", "Membri" o "Members" e "Visitatori" o "Visitors" (a seconda della localizzazione del servizio offerta da Microsoft, i nomi potrebbero essere disponibili anche in altre lingue). I permessi associati a

ciascun ruolo sono riassunti in Tabella 4.1. L'appartenenza di un utente a uno dei tre

Nome Ruolo	Livello di Autorizzazione	Azioni permesse
Proprietari	Controllo completo	Modifica delle impostazioni del sito e del suo contenuto, modifica del livello dei permessi per ciascun gruppo o utente, aggiunta o rimozione di Membri e Visitatori.
Membri	Contributo	Modifica del contenuto del sito, ma non delle impostazioni dello stesso.
Visitatori	Lettura	Sola visualizzazione del contenuto del sito. Nessun tipo di modifica al sito stesso o al suo contenuto.

Tabella 4.1. Ruoli SharePoint e relative autorizzazioni

gruppi determina, dunque, in che modo può accedere a un sito e che tipo di azioni può intraprendere sul sito stesso e sui suoi contenuti. I Proprietari hanno il controllo completo del sito, sono dunque gli unici che possono anche modificare i livelli di accesso al sito e aggiungere o rimuovere utenti dai vari gruppi. Queste autorizzazioni conferiscono loro un potere decisionale completo sul sito, per cui un numero elevato di Proprietari per sito potrebbe rappresentare un segnale di allarme per il rispetto del principio del privilegio minimo. Inoltre, se un sito rimanesse senza alcun utente nel gruppo dei Proprietari, nessun altro all'interno del sito avrebbe l'autorizzazione a inserire o rimuovere utenti, creando una situazione potenzialmente rischiosa per il mantenimento della disponibilità dei dati.

Anche un numero elevato di Membri non giustificato potrebbe compromettere l'integrità dei dati, se tra i Membri ci sono più utenti di quelli che dovrebbero effettivamente avere accesso alla modifica del contenuto del sito.

Oltre ai tre gruppi standard di autorizzazioni, in SharePoint esiste l'opzione di creare gruppi personalizzati per soddisfare esigenze specifiche non coperte dai gruppi predefiniti. Questa funzionalità può essere vantaggiosa poiché permette di creare nuovi gruppi con nomi che rispecchiano meglio la tipologia di utenti che ne fanno parte e con livelli di permesso che si adattano al meglio alle necessità dei team che collaborano su un determinato sito. Tuttavia, la creazione di gruppi personalizzati senza una valida motivazione potrebbe rendere inutilmente complessa la gestione delle autorizzazioni. Sebbene la personalizzazione offra il vantaggio della flessibilità, richiede una gestione particolarmente attenta. Infatti, deviare dall'uso standard, che è definito e più facilmente controllabile, potrebbe introdurre potenziali complicazioni.

Queste sono alcune delle ragioni che hanno portato alla creazione di uno script per raccogliere dati su tutti i siti SharePoint aziendali in modo automatico. Nella Figura 4.4, è illustrato il processo che potrebbe essere effettuato periodicamente, rieseguendo lo script ogni 6 mesi per ottenere una panoramica aggiornata della situazione. Nel diagramma sono

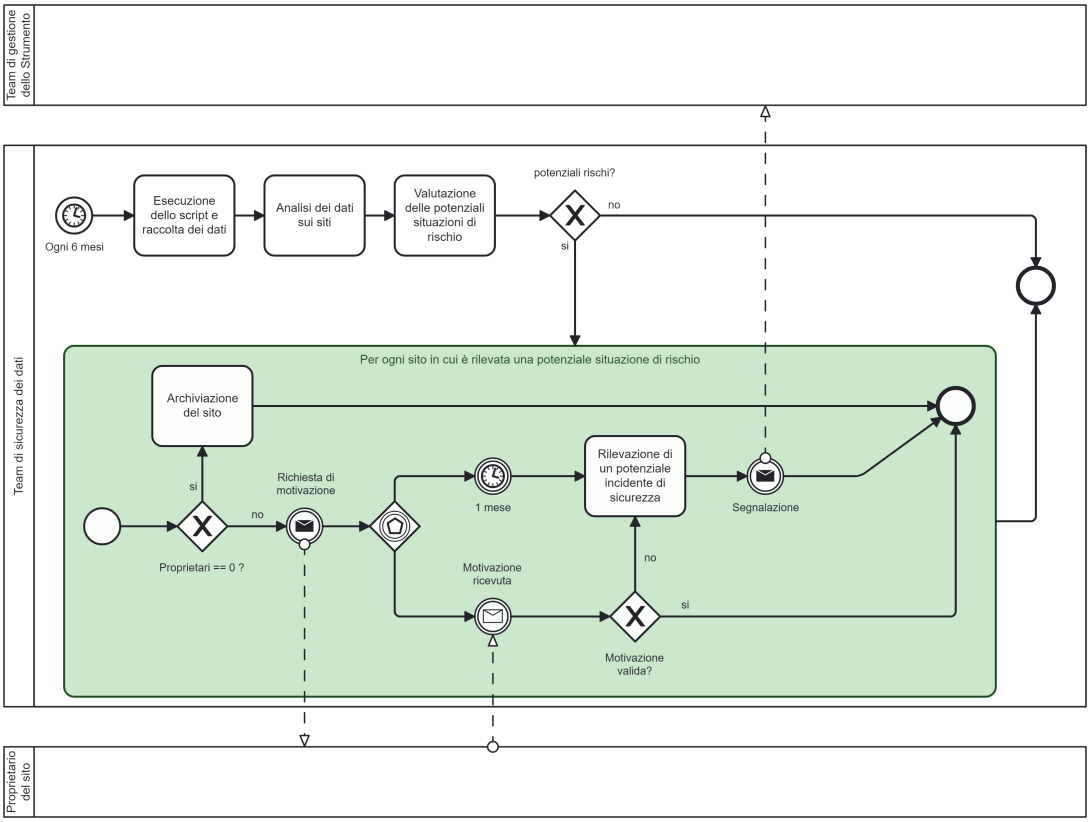

Figura 4.4. Diagramma in BPMN del processo di controllo degli accessi

presenti tre Pool che rappresentano i tre partecipanti al processo: il team di sicurezza dei dati al centro che rappresenta l'organo che svolge le azioni implementate in questo lavoro di tesi, il proprietario del sito SharePoint e un team che si occupa della corretta gestione dello strumento cloud.

Nel diagramma, sono rappresentati tre Pool che simboleggiano i tre attori coinvolti nel processo:

1. In alto, il team che si occupa della corretta gestione delle risorse cloud.
2. Il team di sicurezza dei dati, posizionato al centro, che esegue le azioni delineate in questa tesi.
3. Al fondo, il proprietario del sito SharePoint.

Il processo inizia con l'esecuzione di uno script per la raccolta automatica dei dati sui siti SharePoint, seguita da un'analisi dei dati raccolti. Se l'analisi rivela situazioni potenzialmente anomale o preoccupanti che potrebbero rappresentare rischi o indicare una gestione inadeguata dello strumento, si procede con l'indagine di questi casi specifici, eseguendo il sotto processo nel riquadro in verde per ciascun sito.

Se il problema riscontrato è l'assenza di Proprietari per un sito, il sito viene archiviato, ovvero reso disponibile solo in modalità di lettura, per i motivi precedentemente spiegati. Infatti, in questo caso, non ci sarebbe più nessuno tra gli utenti a poter avere un controllo completo del sito e gestire eventuali modifiche delle impostazioni e dei permessi. Se, invece, il sito ha dei Proprietari ma presenta altre potenziali situazioni di rischio, si procede inviando una email ai Proprietari richiedendo una giustificazione che motivi la gestione non standard dei permessi di accesso al sito.

Se il Proprietario del sito fornisce una motivazione valida, cioè una spiegazione della gestione dei permessi coerente con l'utilizzo del sito, la situazione viene considerata risolta poiché giustificata dalle esigenze operative. Se invece il Proprietario non fornisce una motivazione entro un mese, o se la motivazione fornita non è considerata valida, il caso specifico viene segnalato al team responsabile della corretta gestione degli strumenti cloud.

Ulteriori dettagli sull'implementazione di questo controllo verranno discussi nella sezione 5.2.

4.1.3 Processo di definizione di una strategia per la conservazione dei dati

Il periodo di conservazione dei dati può dipendere da diversi fattori: la tipologia delle informazioni, le normative vigenti, le esigenze aziendali e le capacità di archiviazione.

Nel contesto iniziale del lavoro di tesi, l'organizzazione disponeva già di linee guida che specificano i periodi minimi di conservazione in base alla tipologia di dati trattati. Tuttavia, le direttive necessitano di revisioni periodiche ed eventualmente aggiornamenti per garantire la conformità con le leggi e i regolamenti emergenti, nonché per soddisfare le esigenze aziendali in continua evoluzione e per sfruttare al meglio le nuove tecnologie disponibili.

Una volta definite delle politiche considerate valide dai diversi uffici impattati, si è provveduto a tentare di integrarle con un sistema informativo, così da automatizzare una parte del controllo sulla conservazione ed evitare possibili cancellazioni accidentali da parte degli utenti su file che dovevano essere mantenuti per un periodo minimo.

La scelta del sistema da adottare, in coerenza con le soluzioni implementate precedentemente e tenendo conto della disponibilità di prodotti sul mercato, è ricaduta su una funzionalità offerta da Microsoft Purview. Anche in questo caso, come per la classificazione, il funzionamento è basato su etichette da applicare ai file. In questo contesto, però, non ci si avvale di etichette di riservatezza, ma di etichette di conservazione. Un'etichetta

di conservazione è un'etichetta che si applica ad un file per specificare il tempo per il quale deve essere conservato e cosa fare con esso alla scadenza del periodo di conservazione [23].

Il flusso operativo da seguire è illustrato nella Figura 4.5. Come si può osservare, i

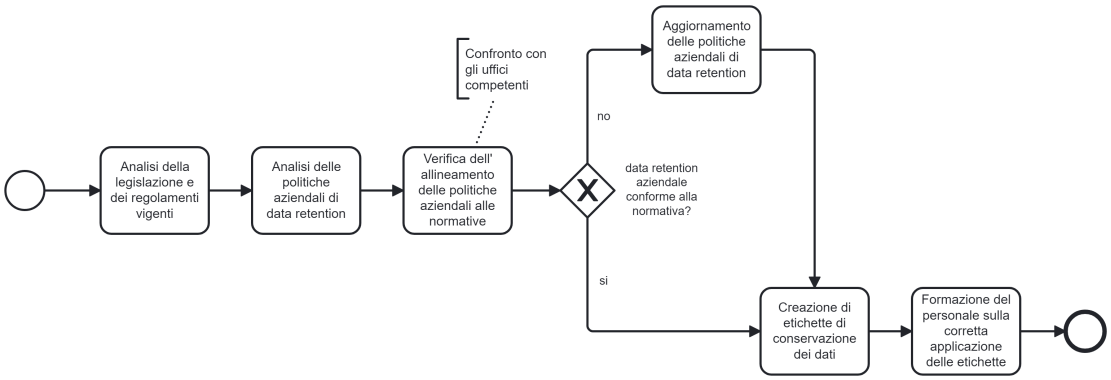

Figura 4.5. Diagramma in BPMN del processo di data retention

primi passaggi consistono nell'analisi dettagliata delle leggi e dei regolamenti vigenti e delle politiche aziendali esistenti per la gestione dei periodi di conservazione dei dati. Successivamente, si verifica l'adeguatezza dell'associazione tra le tipologie di dati e i periodi minimi per ciascuna di esse indicati nelle direttive aziendali. Nel caso di incongruenze, è necessario procedere con un aggiornamento delle direttive.

I passaggi di verifica ed eventuale aggiornamento delle policy devono essere eseguiti in stretta collaborazione con gli uffici competenti. Il loro supporto è fondamentale per garantire che le procedure siano eseguite correttamente, dal momento che questi uffici possiedono una conoscenza approfondita delle normative vigenti e delle esigenze specifiche dell'organizzazione.

Se l'analisi rivela che non sono necessarie modifiche o una volta completato l'aggiornamento, si procede con la creazione di etichette di conservazione. Infine, per garantire una corretta applicazione delle stesse, è essenziale la formazione del personale. Gli utenti devono acquisire consapevolezza dello strumento a disposizione e imparare a riconoscere l'etichetta giusta in base alla tipologia di dato creato.

Capitolo 5

Implementazione

In questo capitolo, verrà descritto nel dettaglio come sono state messe in atto le strategie di protezione dei dati introdotte nel capitolo precedente.

Saranno illustrati: le scelte e i test condotti per configurare le regole di etichettatura automatica, il codice utilizzato per estrapolare i dati relativi agli accessi all'ambiente cloud e la strategia adottata per determinare i periodi di conservazione dei dati.

5.1 Classificazione automatica di dati non strutturati

Come anticipato nel capitolo 4.1.1, l'attività di classificazione dei dati ospitati sui siti SharePoint Online dell'organizzazione è stata eseguita tramite lo strumento Microsoft Purview Information Protection ed è stata realizzata attraverso l'uso di etichette di riservatezza applicate automaticamente in base al contenuto dei file esaminati.

Le etichette possono essere personalizzate in diversi gradi di riservatezza, rispondendo a specifiche necessità all'interno dell'organizzazione. A ciascuna etichetta possono essere assegnate diverse impostazioni di protezione, come la crittografia dei file per prevenire accessi non autorizzati o la definizione del tipo di link di condivisione predefinito, in modo da limitare così la circolazione di documenti riservati. Un'altra opzione è quella di contrassegnare il contenuto per mezzo di intestazioni o filigrane applicate al documento.

In questo lavoro, si è scelto di non applicare restrizioni ai file etichettati in modo automatico per evitare di interferire con le operazioni aziendali. Si ritiene, infatti, che solo l'utente che crea o interagisce direttamente con un documento possa conoscere l'effettiva riservatezza di tale documento e determinarne il livello di protezione più appropriato. Se un documento fosse erroneamente classificato e crittografato, questo potrebbe impedire l'accesso a dipendenti che ne hanno bisogno, alterando o interrompendo le normali operazioni quotidiane. Pertanto, si è preferito che le restrizioni, come la crittografia con chiave simmetrica, potessero essere applicate esclusivamente dagli utenti per via manuale.

Lo scopo di questa fase del lavoro è, dunque, principalmente di fornire un supporto all'utente, dando indicazione sulla natura del contenuto dei documenti qualora questi non fossero già stati classificati manualmente. Ciò permette all'utente di avere una maggiore percezione di come utilizzare, condividere ed eventualmente proteggere adeguatamente il documento. Inoltre, l'importanza della classificazione automatica risiede nella sua capacità di assegnare etichette alla totalità dei documenti supportati in tempi ridotti e con poco sforzo. Queste etichette possono essere facilmente integrate con i sistemi di prevenzione della perdita di dati (DLP) aziendali, contribuendo a prevenire la divulgazione di informazioni riservate all'esterno dell'organizzazione. Infine, avendo un maggior numero di file correttamente classificati, si possono ottenere più facilmente raggruppamenti di file simili per tipologia e che possano necessitare di modalità di protezione affini.

La rilevazione delle informazioni sensibili all'interno dei documenti è stata condotta utilizzando i Tipi di Informazioni Sensibili o Sensitive Information Type (SIT) forniti dallo strumento di classificazione. Una funzionalità alternativa dello strumento Microsoft prevede l'uso del Machine Learning per allenare dei classificatori personalizzati sulle specifiche esigenze aziendali e sulle tipologie di file più frequentemente utilizzate nell'organizzazione. Questi consentirebbero potenzialmente di identificare la riservatezza di una gamma più ampia di informazioni. Tuttavia, il loro impiego apre a scenari di condivisione di file anche sensibili con il fornitore di terze parti (che necessitano di essere regolamentati nei contratti di vendita) e di un uso massiccio della rete per l'addestramento e l'analisi a tempo d'esecuzione dei file stessi.

5.1.1 Configurazione delle etichette

Nel contesto di questa tesi, sono state utilizzate quattro etichette di classificazione già disponibili a cui verrà fatto riferimento con la nomenclatura suggerita da Microsoft: "Pubblico", "Generale", "Riservato", "Estremamente Riservato".

A ciascuna delle etichette è stata assegnata una priorità rappresentata da un numero, in questo caso da 0 a 3. Un numero più basso indica una priorità inferiore e corrisponde a etichette che rappresentano contenuti con un livello di riservatezza più basso. Ad ogni documento può essere applicata una sola etichetta e assicurarsi di assegnare la giusta priorità alle etichette è essenziale per il corretto funzionamento della classificazione automatica. Infatti, se sono presenti più elementi sensibili all'interno del file che corrispondono a più di una etichetta, in automatico viene assegnata al documento quella a priorità più alta, cioè quella che rappresenta il contenuto più riservato all'interno del documento.

È importante sottolineare un principio su cui si basa la classificazione automatica adottata in questo lavoro: il creatore del dato, che ne conosce appieno le finalità, deve poter avere sempre il controllo finale sulla sua categorizzazione. Di conseguenza, il processo di etichettatura automatica non può sovrascrivere una decisione presa manualmente dall'utente, neppure se il classificatore rileva la necessità di applicare un'etichetta con

priorità superiore. Inoltre, l'utente ha in qualsiasi momento la possibilità di modificare il livello di sensibilità assegnato al file, cambiando manualmente l'etichetta impostata automaticamente. Il classificatore automatico può, invece, rietichettare un file precedentemente classificato automaticamente se il contenuto del file cambia nel tempo, sostituendo l'etichetta precedente con una di priorità superiore.

La Tabella 5.1 presenta l'elenco delle etichette utilizzate. Per ciascuna etichetta, si illustra il valore della priorità, i tipi di contenuto corrispondenti e alcuni esempi di documenti che potrebbero rientrare nella categoria.

Nome Etichetta	Priorità	Contenuto del documento	Esempi di documento
Pubblico	0	Dati senza restrizioni destinati alla distribuzione al pubblico.	Comunicati stampa e annunci, fascicoli di prodotti o servizi offerti, resoconti annuali di attività, ecc.
Generale	1	Dati utilizzati quotidianamente e che possono essere condivisi liberamente in tutta l'organizzazione.	Politiche interne e procedure, calendari di riunioni o eventi, documenti di progetto che non contengono dati sensibili, ecc.
Riservato	2	Dati contenenti informazioni personali di dipendenti o clienti e dati cruciali per il raggiungimento degli obiettivi dell'organizzazione.	Piani di progetto interni, strategie di marketing, Documenti con dati di contatto di clienti ecc.
Estremamente Riservato	3	Dati personali sensibili di clienti o dipendenti e, in generale, i dati più critici per l'organizzazione. Possono essere condivisi solo con destinatari specifici.	Informazioni su scelte strategiche, informazioni legali riservate, ecc.

Tabella 5.1. Etichette di riservatezza: priorità e contenuto del file

Le etichette applicabili sono visibili nelle applicazioni che supportano il loro utilizzo e l'utente ha la possibilità di selezionare manualmente quella che ritiene più appropriata per il contenuto del file. Quando si crea un nuovo file o si apre un file esistente, se l'utente non seleziona un'etichetta specifica, l'etichetta "Generale" viene applicata automaticamente come impostazione predefinita, ma potrà essere modificata in qualsiasi momento. L'etichetta applicata con questa modalità viene trattata come un'etichetta applicata dal sistema e, di conseguenza, può essere modificata dal classificatore automatico.

5.1.2 Formati supportati

Microsoft Purview Information Protection consente di individuare delle particolari tipologie di dati sensibili all'interno di documenti di tipo testuale. Si tratta dunque di dati non strutturati, come è stato spiegato nel capitolo 2.1.2.

La possibilità di applicare delle etichette in modo automatico è offerta dallo strumento solo per quattro formati di file:

1. File in Portable Document Format (PDF) con estensione "pdf"
2. File di Excel con estensione "xlsx"
3. File di Word con estensione "docx"
4. File di PowerPoint con estensione "pptx"

Come anticipato alla fine della sezione precedente, se un utente non seleziona un'etichetta manualmente, i file vengono automaticamente classificati dal client con l'etichetta predefinita "Generale" al momento della creazione. Questo processo avviene anche all'apertura di un file già esistente e non ancora etichettato, come nel caso di file che non sono stati aperti da un lungo periodo di tempo, ovvero prima che questa impostazione venisse attivata. Lo stesso vale per i documenti ricevuti da fonti esterne e non ancora aperti sui dispositivi aziendali, poiché al di fuori dell'organizzazione potrebbero essere in vigore politiche di classificazione differenti o potrebbero non esserci affatto.

Tuttavia, è importante sottolineare che questa azione automatica avviene solo per documenti creati o aperti all'interno delle applicazioni di Office, di conseguenza non si applica ai file PDF. Ciò significa che ai file PDF non viene automaticamente assegnata l'etichetta predefinita "Generale" e risultano non classificati, a meno che l'utente non selezioni una specifica etichetta. Questo aspetto sarà rilevante per l'analisi dei risultati, dal momento che per tale formato l'utilizzo dello strumento di classificazione automatica potrebbe risultare particolarmente vantaggioso.

Anche per altri file non Office, come immagini in formato "jpg" o "png", è supportata solo la classificazione manuale da parte dell'utente. Dunque, lo strumento di etichettatura automatica è stato utilizzato con la finalità di coprire una buona parte dei dati aziendali, rappresentati da documenti di tipo testuale, non comprendendo le immagini, che possono ugualmente contenere dati riservati.

5.1.3 Tipi di Informazioni Sensibili

Per comprendere meglio la costituzione delle policy che verranno descritte nelle prossime sezioni del capitolo, è utile chiarire il concetto di "Tipo di Informazioni Sensibili" o "Sensitive Information Type" (SIT) disponibili e come funziona la loro individuazione all'interno dei file.

Come spiegato nella guida ufficiale dello strumento [24], i Sensitive Information Type sono classificatori basati su pattern che rilevano informazioni sensibili e ciascuna entità di tipo SIT è composta dai campi seguenti:

- Nome
- Descrizione
- Modello, costituito a sua volta da:
  - Elemento primario, che può essere un'espressione regolare, un elenco di parole chiave, un dizionario di parole chiave o una funzione.
  - Elemento di supporto, ossia un elemento che contribuisce ad aumentare la confidenza della corrispondenza.
  - Livello di attendibilità, che riflette la quantità di prove di supporto rilevate oltre all'elemento primario.
  - Prossimità, vale a dire il numero di caratteri tra gli elementi primari e quelli di supporto.

In Figura 5.1, è mostrato un documento di esempio contenente i dati di una carta di credito, che sono rilevabili dal SIT chiamato "Numero di carta di credito". L'elemento primario, in questo caso, per essere rilevato deve corrispondere a un numero di cifre che va da 14 a 19 e che superano il test Luhn[1]. Elementi di supporto, invece, possono essere parole chiave situate più o meno vicino all'elemento primario come, ad esempio, una data in un formato valido o parole tipo "carta di credito", "numero di carta".

La probabilità di identificare correttamente un determinato SIT cresce all'aumentare del numero di elementi di supporto che si trovano in prossimità di un elemento primario. Questo indice di confidenza è descritto dal campo chiamato "livello di attendibilità". Ci possono essere tre livelli: basso, medio ed elevato. La Tabella 5.2 illustra i tre livelli, ciascuno con le sue caratteristiche in termini di valore di attendibilità e bilanciamento tra falsi positivi e falsi negativi attesi. Il valore di attendibilità è un numero compreso tra 1 e 100 e descrive l'accuratezza.

5.1.4 Creazione delle policy e test

In questa fase del progetto, è stata definita una policy di etichettatura automatica per ciascuna delle etichette disponibili, ad eccezione dell'etichetta "Pubblico". Questo perché gli strumenti a disposizione, consentono solo la ricerca di informazioni sensibili, non

[1]Il test o formula di Luhn, anche conosciuto come Modulo 10, è un semplice algoritmo che consente di generare e verificare la validità di vari numeri identificativi (numero della carta di credito, codice IMEI, ecc.).

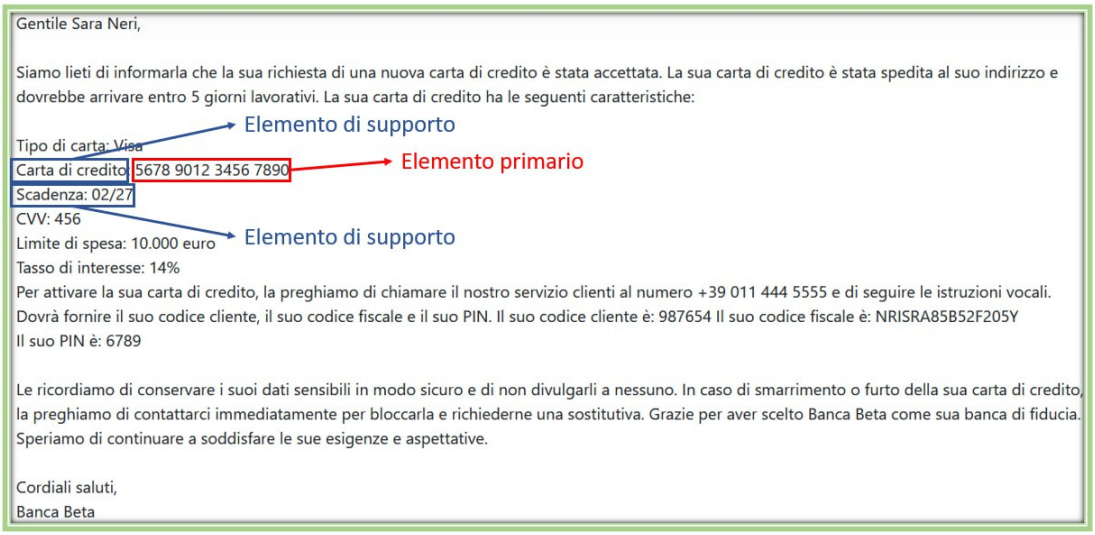

Figura 5.1. Documento di esempio per l'identificazione di un'entità SIT

Livello di attendibilità	FP / FN	Valore di attendibilità (X)
Basso	Restituisce un gran numero di falsi positivi, ma pochi o zero falsi negativi.	$X <= 65$
Medio	Restituisce più falsi positivi che falsi negativi.	$66 < X <= 75$
Elevato	Restituisce il minor numero di falsi positivi, ma potrebbe causare più falsi negativi.	$X > 76$

Tabella 5.2. Livelli di attendibilità dei SIT

permettendo di determinare quando un documento contiene informazioni di pubblico dominio.

In base alla tipologia di documenti, dei dati contenuti al loro interno e del loro grado di sensibilità sono state create tre policy di etichettatura automatica: una policy per i file contenenti dati estremamente riservati, una policy per i file contenenti dati riservati e un'ultima per tutti gli altri file il cui contenuto non rientra in nessuna delle due precedenti regole.

Il funzionamento dello strumento prevede che i file nei formati supportati vengano analizzati per individuare le condizioni specificate nei criteri di classificazione. Ogni qualvolta viene trovata una corrispondenza, viene applicata al file un'etichetta secondo quanto

stabilito dalla policy.

Le tre policy create verranno descritte di seguito, da quella a priorità più alta a quella a priorità più bassa.

Policy per dati estremamente riservati

La policy per l'etichettatura di file contenenti dati estremamente riservati è stata chiamata "Dati estremamente riservati SharePoint".

Questa policy è stata creata sfruttando i SIT pre-configurati disponibili nello strumento. Tra i SIT disponibili, sono stati selezionati quelli che identificano le informazioni più sensibili, quelle la cui diffusione a utenti non autorizzati causerebbe gravi danni a individui o all'organizzazione stessa. Queste informazioni sono quelle che richiedono il più alto livello di controllo e sicurezza e dovrebbero essere accessibili solo ad un numero ristretto di persone. I SIT individuati sono stati raggruppati per categoria e per ciascuna categoria è stata creata una "regola". Ciascuna regola rappresenta un criterio specifico che lo strumento utilizza per applicare l'etichetta. La regola determina quali contenuti e in quale quantità il file deve includere per essere considerato da classificare con una specifica etichetta.

Di seguito è presentata una possibile configurazione delle regole stabilite per questa policy. L'elenco include i nomi delle regole create con alcuni esempi delle corrispondenti entità di tipo SIT:

- **Dati finanziari.** I SIT selezionati sono:
  - Numero di carta di credito
  - Numero di carta di debito dell'Unione europea
- **Informazioni sanitarie.** Questa regola identifica la presenza simultanea di termini legati al campo medico e di identificativi personali. Un esempio dei SIT selezionati è:
  - Numero di patente di guida dell'UE
  - Numero di passaporto UE
  - Numero di identificazione nazionale UE

  AND
  - Classificazione internazionale delle malattie
- **Identificativi medici.** Alcuni esempi di SIT selezionati sono:
  - Numero della Drug Enforcement Agency (DEA)
  - Carta MBI (Medicare Beneficiary Identifier)

  - Codice sanitario personale canadese (PHIN)

- **Credenziali.** Il SIT selezionato è:

  - Tutte le credenziali

Come si evince dall'elenco puntato, la policy applica in automatico l'etichetta "Estremamente riservato" ai file che contengono: numeri di carta di credito o debito, termini sanitari associati a identificativi personali, codici identificativi di servizi sanitari e credenziali di accesso.

È fondamentale sottolineare che nella categoria dei dati estremamente riservati dovrebbero rientrare anche quelli che il Regolamento Generale sulla Protezione dei Dati (GDPR) definisce come dati sensibili. Questi includono, ad esempio, dati relativi all'origine razziale o etnica, alle opinioni politiche, alle convinzioni religiose o filosofiche, all'appartenenza sindacale, nonché dati genetici, dati biometrici intesi a identificare in modo univoco una persona fisica, dati relativi all'orientamento sessuale di una persona [8].

Un discorso analogo vale per informazioni aziendali strategiche che, se divulgate a individui non autorizzati, potrebbero rappresentare una minaccia per la stabilità dell'organizzazione. Tali dati possono includere informazioni relative alla sicurezza informatica, come i risultati delle valutazioni di vulnerabilità o dei test di penetrazione, così come dati finanziari e proprietà intellettuale. Tuttavia, queste informazioni non sono state considerate nella creazione delle regole perché difficilmente rintracciabili tramite espressioni regolari di media o alta complessità, a causa della loro natura imprevedibile (essendo spesso frutto dell'ingegno umano).

Policy per dati riservati

Come per la policy precedente, anche quella per l'etichettatura di file contenenti dati riservati è stata creata sfruttando le SIT.

Questa policy è stata chiamata "Dati riservati SharePoint" e applica in modo automatico l'etichetta di classificazione "Riservato" ai file che soddisfano i criteri stabiliti.

Tale etichetta dovrebbe essere applicata ai file che contengono informazioni il cui accesso deve essere ristretto a un numero limitato di persone all'interno dell'organizzazione e la cui diffusione a utenti non autorizzati potrebbe causare danni a individui o all'organizzazione stessa. Tra questi dati, ci sono quelli che il Regolamento Generale sulla Protezione dei Dati definisce come "dati personali", ossia qualsiasi informazione riguardante una persona fisica identificata o identificabile [8]. Tra questi, dunque, ci sono ad esempio nome e cognome, indirizzo email, indirizzo di abitazione, ma anche indirizzi IP. Similmente alla policy descritta in precedenza, anche in questo caso i SIT preimpostati non sono in grado di coprire tutte le possibili situazioni.

In sintesi, questa policy è stata creata in modo da applicare in automatico l'etichetta "Riservato" a tutti i file contenenti: indirizzi IP, identificativi personali, numeri di conto bancario, nomi di persona e indirizzi fisici.

Di seguito viene presentato l'elenco dei nomi delle regole stabilite per questa policy con alcuni esempi delle corrispondenti entità SIT:

- **Indirizzi IP.** I SIT selezionati sono:
  - Indirizzo IP v4
  - Indirizzo IP v6
- **Identificativi personali.** Alcuni esempi di SIT selezionati sono:
  - Numero di patente di guida dell'UE
  - Numero di passaporto UE
  - Numero di previdenza sociale (SSN) o equivalente UE
  - Codice identificativo del singolo contribuente statunitense (ITIN)
  - Codice PAN (Permanent Account Number) indiano
- **Numeri di conto bancario.** Alcuni esempi di SIT selezionati sono:
  - Numero di conto bancario internazionale (IBAN)
  - Numero di conto bancario degli Stati Uniti
  - Numero di conto bancario australiano
- **Nomi e indirizzi.** I SIT selezionati sono:
  - Tutti i nomi completi
  - Tutti gli indirizzi fisici

Policy per dati generali

L'ultima policy, è stata creata per l'etichettatura dei dati che non rientrano nelle categorie precedentemente menzionate con l'etichetta di classificazione "Generale". Questi dati sono quelli che possono liberamente circolare all'interno del perimetro aziendale come nel caso, ad esempio, di regolamenti interni, ma la cui diffusione all'esterno potrebbe rappresentare un rischio, se pur medio-basso, per l'organizzazione.

La creazione di questa policy non è stata effettuata mediante l'uso di entità SIT, poichè queste identificano solo dati con un determinato livello di sensibilità che i dati da considerare come "generali" non hanno. Invece, la policy è stata formulata con un'unica regola che si basa sul formato del file. Nella regola è stato indicato che tutti i file trovati con estensione pdf, xlsx, docx o pptx sono da classificare con l'etichetta "Generale". Di conseguenza, la policy prende in considerazione tutti i formati che la tecnologia utilizzata può etichettare in modo automatico, assegnando ai file corrispondenti l'etichetta con priorità 1.

In base al principio spiegato in 5.1.1, tra tutti i file analizzati dallo strumento, se sono già state identificate corrispondenze con una delle due etichette a più alta priorità, questa non potrà sovrascriverle. Il risultato finale sarà, dunque, che solo i file che non soddisfano i criteri delle altre due policy di etichettatura verranno contrassegnati con l'etichetta "Generale".

É stata presa la scelta implementativa di classificare in modo automatico tutti i documenti con al minimo l'etichetta "Generale", assumendo tutti i documenti come destinati esclusivamente all'uso e alla condivisione libera esclusivamente all'interno dell'organizzazione. Ciò significa che anche i documenti pubblici, come i documenti scaricati dal Web o quelli contenenti informazioni aziendali di pubblico dominio, sono stati considerati come destinati all'uso interno. Nonostante questa tipologia di documenti potrebbe essere diffusa senza pericoli anche fuori dall'organizzazione, distinguere automaticamente questa particolare casistica, basandosi solo sul contenuto del documento non è possibile con gli strumenti utilizzati.

L'assunzione di base è stata, dunque, quella di considerare tutti i file che vengono salvati sui siti SharePoint come utilizzabili liberamente solo all'interno dell'azienda, a meno che l'utente non provveda manualmente a etichettarli come "Pubblico".

Test sulle policy

Prima di raggiungere la configurazione definitiva delle policy, sono state effettuate diverse simulazioni su alcuni siti SharePoint di cui era noto il contenuto di tutti i file presenti. Dai risultati ottenuti e dalla valutazione della presenza di più o meno falsi positivi e falsi negativi, le policy sono state affinate in relazione alle tipologie di documenti che circolano nell'azienda.

Per quanto riguarda la policy "Dati generali SharePoint", considerata la sua semplicità e il fatto che non si basa sui tipi di informazioni sensibili, è stata fatta un'unica simulazione che ne verificasse il corretto funzionamento.

Le altre due policy, al contrario, hanno subito alcune modifiche per quanto riguarda la configurazione delle regole, il settaggio del livello di attendibilità e del contatore di occorrenze dei SIT.

Inizialmente, il livello di attendibilità di tutti i SIT utilizzati è stato impostato secondo le raccomandazioni dello strumento. Tuttavia, dopo alcune simulazioni, si è deciso di impostare ciascun SIT a un livello di attendibilità elevato, al fine di minimizzare la possibilità di falsi positivi.

Di seguito, si fornisce un elenco delle modifiche principali che sono state identificate come necessarie per ottimizzare i risultati forniti dalle policy nell'ambiente di test:

Modifiche alla policy "Dati riservati SharePoint"

1. É stato notato che i SIT "Tutti i nomi completi" e "Tutti gli indirizzi fisici" presi singolarmente restituivano diversi falsi positivi. Per questa ragione, si è provato

ad impostare il criterio di ricerca con l'operatore logico AND, stabilendo che per soddisfare la regola un file dovesse presentare almeno uno dei due SIT.

In questa circostanza, nonostante fossero scomparsi i casi di falsi positivi, erano notevolmente aumentati i falsi negativi. Infatti, non venivano rilevati alcuni nomi e indirizzi in file in cui i SIT erano presenti singolarmente, l'uno senza l'altro.

La decisione finale è stata quella di ripristinare la ricerca dei due SIT come indipendenti l'uno dall'altro, ma aumentando il contatore delle occorrenze di nomi e indirizzi da ricercare in un file. Il contatore è stato settato a minimo 20 occorrenze. Questo compromesso, implica che la regola limita l'applicazione dell'etichetta solo ai file che contengono elenchi di nomi o di indirizzi.

La presenza di liste di nomi e indirizzi aumenta la probabilità che si tratti di documenti effettivamente confidenziali, come quelli che contengono dati di dipendenti o clienti, rispetto a documenti non riservati in cui compare semplicemente una firma, il nome dell'autore di una presentazione o l'indirizzo delle sedi aziendali. In questo modo, si evita di classificare come riservati documenti che necessitano di essere diffusi a livello generale in tutta l'organizzazione.

Modifiche alla policy "Dati estremamente riservati SharePoint"

1. Nella regola "Informazioni sanitarie", oltre al SIT "Classificazione internazionale delle malattie", sono stati inizialmente inseriti altri SIT riguardanti l'ambito sanitario come "Termini di analisi di laboratorio" e "Specializzazioni mediche". Questi SIT hanno generato dei falsi positivi nei documenti testati. Poiché il settore sanitario non è strettamente pertinente all'ambito aziendale in questione, il SIT è stato ridotto al minimo per limitare i falsi positivi.
2. Nella regola "Credenziali", il contatore delle occorrenze del SIT era inizialmente impostato da 1 a qualsiasi. Questa impostazione ha generato diversi falsi positivi in documenti aziendali in cui, ad esempio, veniva spiegato come impostare una password o venivano fornite credenziali false a scopo illustrativo. Pertanto, la regola finale è stata modificata impostando il contatore delle occorrenze a un minimo di 2. Questo significa che un file soddisfa la regola se contiene da 2 a un numero qualsiasi di credenziali utente.

A seguito di queste modifiche è stato raggiunto lo stato definitivo delle policy, che viene riassunto nella sezione successiva.

5.1.5 Configurazione definitiva delle policy

In seguito alle modifiche apportate alle policy durante la fase di simulazione, si è giunti alla configurazione finale illustrata in Figura 5.2.

Nome della policy	Descrizione	Etichetta da applicare	Nome regola	Configurazione della regola e SIT	Livello di attendibilità	Occorrenza dei SIT
Dati generali SharePoint	Questa policy applica automaticamente l'etichetta "Generale" ai file, nei siti Sharepoint, che contengono file con tutte le estensioni supportate dall'etichettatura automatica: pdf, xlsx, docx e pptx.	Generale	Tutti i formati	L'estensione del file allegato è: • pdf • xlsx • docx • pptx		
Dati riservati SharePoint	Questa policy applica automaticamente l'etichetta "Riservato" ai file, nei siti Sharepoint, che contengono: indirizzi IP, identificativi personali, numeri di conti bancari, nomi e cognomi, indirizzi fisici.	Riservato	Indirizzi IP	Il file contiene uno qualsiasi dei seguenti elementi: • Indirizzo IPv4 • Indirizzo IPv6	Elevato	1-qualsiasi
			Identificativi personali	Il file contiene uno qualsiasi dei seguenti elementi: • Numero di patente di guida dell'UE • Numero di passaporto UE • SSN o equivalente UE • Codice ITIN • Codice PAN indiano • Ecc.	Elevato	1-qualsiasi
			Numeri di conto bancario	Il file contiene uno qualsiasi dei seguenti elementi: • IBAN • Numero conto americano • Numero conto australiano • Ecc.	Elevato	1-qualsiasi
			Nomi e indirizzi	Il file contiene uno qualsiasi dei seguenti elementi: • Tutti i nomi completi • Tutti gli indirizzi fisici	Elevato	20-qualsiasi
Dati estremamente riservati SharePoint	Questa policy applica automaticamente l'etichetta "Estremamente Riservato" ai file, nei siti SharePoint, che contengono: numeri di carte di credito o di debito, termini collegati a malattie associate a indentificativi personali, identificativi medici, credenziali.	Estremamente Riservato	Dati finanziari	Il file contiene uno qualsiasi dei seguenti elementi: • Numero di carta di credito • Numero di carta di debito UE	Elevato	1-qualsiasi
			Informazioni sanitarie	Il file contiene uno qualsiasi dei seguenti elementi: • Numero di patente di guida UE • Numero di passaporto UE • Numero identificazione nazionale UE • Ecc. **AND** Il file contiene uno qualsiasi dei seguenti elementi: • Classificazione internazionale delle malattie	Elevato	1-qualsiasi
			Identificativi medici	Il file contiene uno qualsiasi dei seguenti elementi: • Numero della DEA • Carta MBI • PHIN • Ecc.	Elevato	1-qualsiasi
			Credenziali	Il file contiene uno qualsiasi dei seguenti elementi: • Tutti i tipi di credenziali	Elevato	2-qualsiasi

Figura 5.2. Configurazione definitiva delle policy di etichettatura automatica

Nell'immagine sono riportati per ciascuna delle tre policy: il nome, la descrizione, l'etichetta da applicare, le regole e alcuni esempi di SIT selezionati, il livello di attendibilità delle entità SIT selezionate e l'occorrenza dei SIT. In particolare, le ultime due colonne mostrano un valore unico che è stato selezionato in modo identico per tutti i SIT presenti nella regola.

Una volta raggiunta questa configurazione, si è deciso di attivare le policy su tre siti SharePoint Online per valutarne gli effetti, prima di estendere l'applicazione su larga scala. I risultati finali dell'applicazione sui tre siti offrono un'anteprima del comportamento degli utenti e dell'importanza dell'etichettatura automatica. Questi aspetti saranno analizzati più approfonditamente nel prossimo capitolo.

5.2 Controllo degli Accessi

Questa parte del lavoro di tesi è stata dedicata all'estrazione di dati relativi ai permessi di accesso ai siti SharePoint Online utilizzati in azienda.

Come anticipato nel capitolo 3, SharePoint è un servizio cloud per la gestione dei documenti che può essere utilizzato per creare siti che fungano da spazio sicuro per archiviare, organizzare e condividere informazioni. É possibile creare un sito direttamente dalla pagina principale di SharePoint, ma non solo. SharePoint è strettamente integrato con altri strumenti di Microsoft 365 come, ad esempio, Teams [27], che è molto utilizzato nelle organizzazioni per comunicare in tempo reale e condividere documenti. Ogni volta che viene creato un nuovo team in Teams, viene automaticamente creato anche un sito SharePoint associato ad esso, dove vengono archiviati i file che i componenti del team condividono fra loro. In questi casi, il sito SharePoint viene associato a un gruppo Microsoft 365.

Il concetto di gruppo Microsoft 365 facilita l'integrazione dei vari servizi offerti dalla suite di Microsoft 365 e identifica, quindi, un insieme di utenti che possono accedere contemporaneamente a una serie di servizi.

I gruppi Microsoft 365 e i ruoli di SharePoint, illustrati nella Tabella 4.1, sono entrambi utilizzati per la gestione delle autorizzazioni, ma presentano alcune differenze. In ciascun gruppo Microsoft 365 si possono distinguere un gruppo di Proprietari e un gruppo di Membri, mentre il sito SharePoint dispone di tre ruoli standard: Proprietari, Membri e Visitatori.

Quando il sito SharePoint viene generato in seguito alla creazione di un nuovo team su Teams, i ruoli SharePoint vengono gestiti automaticamente utilizzando il gruppo Microsoft 365 collegato. Questo vuol dire che l'aggiunta di Proprietari e Membri al gruppo o al team fa sì che questi vengano aggiunti anche come Proprietari e Membri del sito. In sintesi, ogni gruppo Microsoft 365 ha dei Proprietari e dei Membri che Coincidono con i Proprietari e Membri del team di Teams e con i ruoli di Proprietari e Membri in SharePoint. Il collegamento viene illustrato in Figura 5.3, dove si può anche notare che il ruolo

di Visitatore in SharePoint, non esiste nei gruppi di Microsoft 365. Pertanto, se si desidera aggiungere utenti con autorizzazioni di sola visualizzazione al sito, bisogna aggiungerli direttamente al gruppo Visitatori del sito tramite SharePoint. Questa procedura di ag-

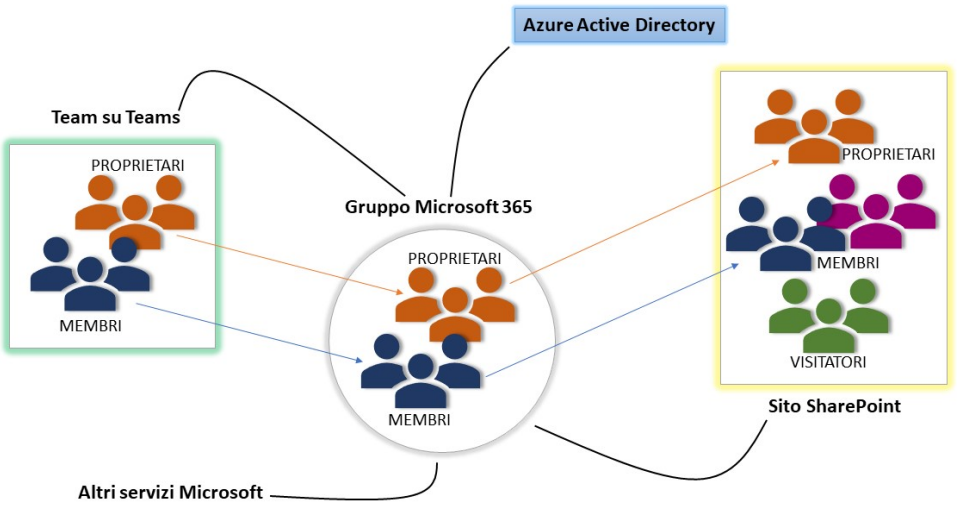

Figura 5.3. Rappresentazione del collegamento tra i gruppi di autorizzazioni in Microsoft 365

giunta di utenti può essere applicata anche per altri ruoli, come mostrato in Figura 5.3 con i Membri colorati in viola. Tuttavia, è importante sottolineare che le autorizzazioni impostate su un sito SharePoint al di fuori del gruppo Microsoft 365 implicano che quegli utenti avranno accesso esclusivamente al sito SharePoint Online e non agli altri servizi a cui può accedere il gruppo.

Un altro concetto correlato ai precedenti è che quando si crea un gruppo Microsoft 365, viene generato un oggetto gruppo in Azure Active Directory (Azure AD). Azure Active Directory è un servizio che garantisce Autenticazione, Autorizzazione e Contabilizzazione degli utenti, regolando gli accessi alle applicazioni aziendali configurate per usarlo [19]. L'oggetto gruppo creato in Azure AD contiene le informazioni di base e i componenti del gruppo Microsoft 365 da cui deriva e possiede un identificativo univoco.

5.2.1 Modelli di siti SharePoint

A seconda del tipo di sito che si decide di creare, SharePoint mette a disposizione vari modelli. Le due opzioni principali che sono state individuate all'interno dell'azienda sono:

- Sito di Team: progettato per facilitare la collaborazione all'interno di un team e corrisponde, dunque, a un sito privato.

- Sito di Comunicazione: rappresenta un sito pubblico utilizzato come portale per diffondere notizie, avvisi e servizi al pubblico aziendale.

A loro volta, questi modelli possono avere dei nomi specifici a seconda di alcune loro proprietà e caratteristiche aggiuntive. Nel corso di questo studio, è stata rilevata la presenza dei modelli di sito più diffusi in azienda:

Modello STS#3 : Un sito di tipo team che non ha connessioni con il gruppo Microsoft 365.

Modello SITEPAGEPUBLISHING#0 : Un sito di tipo comunicazione, non connesso ai gruppi Microsoft 365, ma sfrutta i ruoli classici di SharePoint.

Modello GROUP#0 : Un sito di tipo team con collegamento ai gruppi Microsoft 365. Questo è anche il modello di sito che viene creato in automatico alla creazione di un nuovo team su Teams.

5.2.2 Definizione dei dati da ricavare

Prima di sviluppare lo script per raccogliere i dati sugli accessi, è stato necessario definire quali dati si intendeva cercare. Sono state quindi esaminate alcune funzionalità e caratteristiche che i siti SharePoint possono avere e quali potrebbero essere i potenziali rischi o i comportamenti degli utenti da monitorare.

Numero di utenti per ciascun gruppo standard

Uno degli obiettivi è stato assicurarsi di avere una corretta distribuzione delle autorizzazioni, monitorando il numero di utenti per ciascun gruppo che identifica i ruoli standard di Proprietari, Membri e Visitatori. Infatti, in base alle autorizzazioni concesse a ciascun gruppo, è importante garantire che il numero di utenti che ne fanno parte sia proporzionato al livello di controllo che hanno sul sito. In altre parole, ci si aspetta di avere un numero massimo di Proprietari che non superi una certa soglia, valutata come ragionevole intorno a 100. Questi utenti hanno il controllo totale del sito e non è consigliabile concedere questi permessi a chiunque, piuttosto è preferibile limitarli a un numero ristretto di utenti. Allo stesso tempo, è importante che il gruppo dei Proprietari non rimanga mai vuoto, poiché in tal caso non ci sarebbe più nessuno all'interno del sito in grado di aggiungere o eliminare utenti e modificare le impostazioni del sito stesso. Pertanto, oltre a una soglia massima da non superare, è necessario monitorare anche la presenza di siti senza proprietari e, se necessario, applicare un'azione correttiva. Questa potrebbe consistere nell'archiviare il sito per renderlo in sola lettura o nell'assegnare a almeno due utenti il ruolo di proprietario da parte dell'amministratore del sito. Per i componenti del gruppo Membri, che hanno un minor numero di permessi, il numero di utenti va comunque monitorato, ma la soglia può essere alzata. Si è stabilito come limite massimo il numero di 500 Membri. Il

ruolo dei Visitatori concede permessi di sola lettura, il che significa che un'assegnazione errata di questa autorizzazione a utenti che non dovrebbero averla potrebbe comportare una perdita di riservatezza delle informazioni. In questo caso, considerata la possibilità di un danno potenziale limitato, non sono state stabilite soglie specifiche per i gruppi di Visitatori. Nonostante ciò, è sempre consigliabile monitorare anche questo gruppo e assicurarsi che un numero molto elevato di utenti in lettura sia presente solo in casi di siti di comunicazione, ovvero siti pubblici in cui si desidera che un ampio numero di dipendenti abbia accesso agli avvisi e ai contenuti trasmessi attraverso il sito.

Gruppi di autorizzazioni personalizzati

Tra le funzionalità offerte da SharePoint, come anticipato nel capitolo 4.1.2, c'è la possibilità di creare gruppi di autorizzazioni personalizzate, che differiscono da quelli standard di SharePoint. L'utilizzo di questi gruppi di autorizzazioni può essere particolarmente utile per personalizzare la gestione degli accessi ai siti in base a esigenze specifiche. Tuttavia, la personalizzazione si discosta da un comportamento standard e avere numerosi gruppi personalizzati o numerosi utenti al loro interno può complicare il controllo degli accessi, poiché per ciascuno di questi casi sarebbe necessario verificare la presenza di una valida giustificazione.

I gruppi personalizzati hanno una denominazione diversa da quelli standard e possono quindi essere riconosciuti in base al titolo.

Gruppi speciali

In SharePoint esistono dei gruppi speciali come i gruppi "Everyone" o "Everyone Except External Users" che contengono rispettivamente tutti gli utenti appartenenti al dominio dell'Active Directory o solo quelli interni all'organizzazione.

Nei siti pubblici, l'uso dei gruppi speciali può essere vantaggioso rispetto all'aggiunta manuale di numerosi gruppi Microsoft 365 o di utenti singoli. Il loro utilizzo nei siti privati, invece, può rappresentare un segnale di allarme. La presenza di questi gruppi speciali può essere riscontrata sia nel gruppo dei Membri sia in quello dei Visitatori di un sito. All'interno del gruppo standard Membri, il gruppo speciale concede a chiunque l'autorizzazione di apportare modifiche al contenuto del sito e per questo motivo va opportunamente giustificata.

Collegamenti Condivisibili e Accessi Limitati

Un'altra funzionalità che SharePoint mette a disposizione per accedere ai contenuti di un sito è quella della creazione di link condivisibili o "Shareable Links". Questi collegamenti possono essere utilizzati per condividere un singolo file o una singola cartella senza necessità di concedere l'accesso a tutto il contenuto del sito. Esistono tre tipi di collegamento principali:

1. “Chiunque” o “Everyone” che danno accesso all’elemento a chiunque abbia a disposizione il collegamento, incluse persone esterne all’organizzazione.

2. “Tutti gli utenti dell’organizzazione” che dà accesso a chiunque all’interno dell’organizzazione che abbia il collegamento.

3. “Persone specifiche” che dà accesso solo agli utenti specificati al momento della creazione del collegamento.

Dopo la creazione del link, il collegamento viene copiato e distribuito a seconda delle necessità. L’uso di un link condivisibile facilita la condivisione di un documento ed è particolarmente utile quando si desidera che alcuni utenti collaborino solo su una parte delle risorse del sito. D’altra parte, questo può rappresentare un rischio se le impostazioni non vengono selezionate con attenzione, soprattutto nel caso di link “Everyone”, che potrebbero essere diffusi in modo improprio.

In SharePoint, è possibile utilizzare anche la funzione “Limited Access”. Si tratta di un livello di autorizzazione che permette a un utente o ad un gruppo di accedere a una specifica pagina del sito o a una libreria senza che essi abbiano i permessi per aprire o modificare altri elementi nel sito. Il concetto è simile a quello degli “Shareable Links”, ma in questo caso non si creano collegamenti da copiare e distribuire. Piuttosto, si seleziona direttamente un contenuto, un utente e si definisce con quali permessi può accedervi. Le considerazioni fatte per gli “Shareable Links” sono quindi valide anche in questo caso, ma il rischio è inferiore perché, se viene selezionato un utente specifico, è meno probabile che ci sia una diffusione impropria di informazioni.

Dalle considerazioni fatte sugli Shareable Links e i Limited Access, emerge l’importanza di un utilizzo prudente di un monitoraggio costante dell’uso di queste funzionalità. In caso di utilizzo eccessivo, si è ritenuto necessario chiedere una spiegazione ai Proprietari del sito. L’obiettivo è capire se l’uso di tali collegamenti è veramente indispensabile, oppure se potrebbe essere evitato assegnando direttamente gli utenti ai loro rispettivi ruoli standard.

5.2.3 Sviluppo dello script

É stato sviluppato uno script specifico per l’estrazione dei dati utilizzando PowerShell, una soluzione di automazione multipiattaforma che comprende una shell della riga di comando, un linguaggio di scripting e un framework di gestione della configurazione [22].

La scelta di Powershell, rispetto ad altri linguaggi di scripting, è motivata dalla sua efficienza nell’interazione con una vasta gamma di strumenti Microsoft 365. PowerShell è stato progettato specificamente per lavorare con questi servizi, il che lo rende ideale per l’estrazione di dati ai fini di questo lavoro. Lo strumento supporta molti comandi e funzionalità che consentono di personalizzare l’estrazione dei dati in base alle esigenze specifiche. Questo livello di dettaglio non è facilmente raggiungibile attraverso altre interfacce, come quella dello stesso SharePoint.

Il codice sviluppato per l'estrazione dei dati sugli accessi è stato riformulato in uno pseudocodice. La decisione è stata presa al fine di semplificare la leggibilità e la comprensibilità del codice. Questo approccio consente di focalizzarsi sulla logica dell'algoritmo e sul suo obiettivo, piuttosto che sui dettagli sintattici del codice. Tali dettagli, infatti, sono peculiari del linguaggio utilizzato, mentre la logica impiegata può essere facilmente adattata per essere utilizzata con altri linguaggi di programmazione.

L'algoritmo implementato è quello mostrato in Figura 5.4. Le variabili principali sono le seguenti:

- "data" contiene la lista di siti SharePoint. Questa lista è stata importata da un file XML ottenuto, dopo essersi connessi a SharePoint Online tramite Powershell, utilizzando il comando

```
Get-SPOSite -Limit All
```

- "count_owners", "count_members", "count_visitors", "count_other" indicano rispettivamente il numero di Proprietari, di Membri, di Visitatori e di gruppi personalizzati presenti nel sito.
- "count_LA" e "count_SL" sono rispettivamente i contatori dei Limited Access e degli Shareable Links trovati nel sito.
- "flagAllUser_members", "flagAllUser_visitors", "flagAllUser_others" sono dei flag utilizzati per tenere traccia della presenza del gruppo speciale "Everyone" rispettivamente all'interno dei gruppi Membri, Visitatori e Altri gruppi non standard.

L'algoritmo itera su ciascun sito della lista e, per ciascun sito, sui gruppi appartenenti al sito. Durante questi cicli, l'algoritmo esegue un conteggio degli utenti in ciascun gruppo, basandosi sul titolo del gruppo. Le condizioni per definire il tipo di gruppo su cui si sta iterando si basano sulla presenza di determinate parole chiave nel titolo. Ad esempio, il gruppo dei Proprietari del sito chiamato "Tesi di laurea" avrà il titolo "Tesi di laurea Owners". Allo stesso modo, il gruppo dei Membri avrà il titolo "Tesi di laurea Members", e il gruppo dei Visitatori avrà il titolo "Tesi di Laurea Visitors". Pertanto, per identificare questi gruppi, l'algoritmo distingue i titoli in base alla parola finale, che può essere "Owners", "Members" o "Visitors".

Per quanto riguarda il gruppo dei Proprietari, se la proprietà "roles", che definisce i ruoli per i componenti del gruppo, non è definita, viene impostata una variabile per per segnalare che il conteggio dei Proprietari per il sito in questione dovrà essere azzerato. Questo controllo viene effettuato anche per i gruppi dei Membri e dei Visitatori, verificando che il campo dei ruoli non sia vuoto.

Per i gruppi Membri e Visitatori, l'algoritmo esegue un'iterazione sulla lista $users di utenti che ne fanno parte. Questa lista può includere sia utenti singoli che gruppi di utenti.

```
$data = Lista dei siti di Sharepoint
for $site in $data :
    for $group in $site :
        if $group.Title == "*Owners" :
            if $group.Roles == NULL :
                $nullRoles_flag = 1
        if $group.Title == "*Members" :
            if $group.Roles != NULL :
                for $user in $group.users :
                    if $user == "*@*" : $count_members++
                    else :
                        if $user != "spo-grid-all-users*" :
                            $owners = prendo gli owner del sito
                            $count_owners = conto il numero di proprietari
                            $members = prendo i membri del sito
                            #ciclo da eseguire su tutti i sottogruppi
                            for $member in $members :
                                if $member.ObjectType == "Group" :
                                    $subGroupMembers = prendo i membri del sottogruppo
                                    $members += $subGroupMembers
                                else : $single_members += $member
                            $uniqMembers = rimuovo i duplicati da $single_members
                            $count_members += $uniqMembers.count
                        else : $flagAllUser_members=1
        if $group.Title == "*Visitors" :
            if $group.Roles != NULL :
                for $user in $group.users :
                    if $user == "*@*" : $count_visitors++
                    else :
                        if $user != "spo-grid-all-users*" :
                            $members = prendo i membri di $users
                            #ciclo da eseguire su tutti i sottogruppi
                            for $member in $members :
                                if $member.ObjectType == "Group" :
                                    $subGroupMembers = prendo i membri del sottogruppo
                                    $members += $subGroupMembers
                                else : $single_members += $member
                            $uniqMembers = rimuovo i duplicati da $single_members
                            $count_visitors += $uniqMembers.count
                        else : $flagAllUser_visitors=1
        if $group.Title == "Limited*" :
            if $group.Roles != NULL :
                $count_LA++
        if $group.Title == "Sharing*" :
            if $group.Roles != NULL :
                $count_SL++
        if $group.Title != [Da tutti i Title degli IF sopra] :
            if $group.Roles != NULL :
                for $user in $group.users :
                    if $user == "*@*" : $count_other++
                    else :
                        if $user != "spo-grid-all-users*" :
                            if $user != "SHAREPOINT\system" :
                                $members = prendo i membri di $users
                                #ciclo da eseguire su tutti i sottogruppi
                                for $member in $members :
                                    if $member.ObjectType == "Group" :
                                        $subGroupMembers = prendo i membri del sottogruppo
                                        $members += $subGroupMembers
                                    else : $single_members += $member
                                $uniqMembers = rimuovo i duplicati da $single_members
                                $count_other += $uniqMembers.count
                        else : $flagAllUser_others++
    if $nullRoles_flag == 0 : $count_owners=0
    $count_all = $count_owners + $count_members + $count_visitors + $count_other
    Aggiungo una tupla al CSV finale contenente tutti i dati ricavati sopra
$out = Esporto il CSV
```

Figura 5.4. Pseudocodice per l'estrapolazione dei dati nei siti SharePoint

Nel caso di utenti singoli, questi vengono identificati tramite l'indirizzo email e vengono aggiunti al conteggio, incrementando i contatori $count_members e $count_visitors per i gruppi Membri e Visitatori rispettivamente. Nel caso di gruppi di utenti, può essere presente il gruppo speciale Everyone, identificato dalla ricerca della parola chiave "spogrid-all-users". In questo caso, viene impostato un indicatore per segnare la sua presenza. Tutti gli altri gruppi di utenti sono presenti nella lista con il codice univoco che li identifica in Azure Active Directory. Per estrarre i Membri di questi gruppi si utilizza il comando:

```
Get-AzureADGroupMember -ObjectId $user -All $true
```

dove in $user è memorizzato il codice identificativo del gruppo. Dal momento che è possibile avere gruppi annidati, è necessario iterare su ciascun membro ottenuto da questi comandi. Se si tratta di utenti singoli, questi possono essere aggiunti all'array $single_users. Se invece si tratta di sottogruppi, l'algoritmo deve estrarre ricorsivamente la lista dei membri dal sottogruppo da Azure AD tramite il comando riportato in precedenza. Al termine del ciclo per estrarre tutti gli utenti singoli dai sottogruppi, l'algoritmo filtra gli utenti ottenuti eliminando i duplicati e li aggiunge al contatore. Quando si itera nel gruppo dei Membri, oltre ai membri del sito e con lo stesso codice identificativo del gruppo, è possibile ricavare i Proprietari utilizzando il comando:

```
Get-AzureADGroupOwner -ObjectId $user -All $true
```

il cui risultato viene usato per conteggiare il numero di utenti, aggiungendoli alla variabile $count_owners.

Il conteggio degli Shareable Links e dei Limited Access è stato effettuato cercando le parole chiave "Limited" e "Sharing" nel titolo del gruppo e aggiornando i contatori corrispondenti quando queste parole chiave sono presenti.

Tutti i gruppi il cui nome non contiene tutte le parole chiave menzionate in precedenza sono stati considerati nel conteggio dei gruppi personalizzati. Il conteggio degli utenti che vi appartengono è stato fatto in modo analogo al gruppo dei Visitatori.

Infine, nella variabile $count_all, sono stati sommati il numero di proprietari, di membri, di visitatori e di utenti appartenenti ad altri gruppi per ottenere il numero totale di utenti per sito.

I risultati dello script sono stati esportati in un file CSV per facilitare la lettura e l'analisi dei dati.

5.3 Conservazione dei dati

Nel corso di questa ricerca, sono state gettate le basi per un progetto di revisione e delle politiche di conservazione dei dati all'interno dell'azienda. Tuttavia, è fondamentale

sottolineare che si prevede il proseguimento di questo lavoro oltre la conclusione dello studio, poiché non è stata ancora realizzata alcuna implementazione pratica.

La conservazione dei dati è un tema sensibile, che richiede una particolare attenzione e una stretta collaborazione tra gruppi di lavoro con competenze diverse. La creazione e l'implementazione di politiche efficaci di conservazione dei dati richiedono la comprensione approfondita delle normative vigenti e una conoscenza dettagliata delle esigenze specifiche dell'organizzazione, nonché delle implicazioni di sicurezza.

All'interno dell'azienda, al momento dell'inizio del lavoro, erano già state stabilite delle politiche per definire la durata di conservazione per le varie tipologie di dati trattati. Mancava un sistema applicativo che impedisse in modo automatico la cancellazione accidentale dei dati da parte dell'utente prima del termine stabilito per la conservazione. Questo lavoro ha permesso di sviluppare le fondamenta per definire, nel contesto attuale, l'integrazione di un sistema che impedisca la cancellazione dei dati prima della scadenza del periodo minimo di conservazione. Questo potrà ridurre il rischio di perdita accidentale di informazioni cruciali o che devono rimanere memorizzate per motivi legali.

In questa fase di lavoro, l'attenzione non è stata rivolta alla cancellazione automatica dei file dopo un determinato periodo di tempo, ma alla garanzia che questi vengano conservati per il tempo necessario. La cancellazione automatica di informazioni che devono essere conservate richiede delle riflessioni aggiuntive, dal momento che solo l'utente che crea, utilizza e modifica i dati ha una conoscenza completa dei loro dettagli. Di conseguenza, l'obiettivo è di conservare i dati per il tempo adeguato, garantendo allo stesso tempo che l'utente, nel frattempo, abbia il controllo e la responsabilità sulla loro gestione.

Nel contesto del lavoro di tesi, è stato identificato Microsoft Purview come uno strumento potenzialmente idoneo per gestire la conservazione dei dati dell'organizzazione ospitati sul cloud. Analogamente a quanto considerato per le etichette di classificazione, si ritiene che potrebbe essere utile l'implementazione di etichette di conservazione.

Il modello che si è pensato di implementare prevede che ogni manager formi i propri dipendenti sull'uso corretto delle etichette di conservazione. Gli utenti avranno a disposizione delle etichette preimpostate e riceveranno istruzioni sulle etichette che sono più adatte per i tipi di dati con cui lavorano abitualmente nel loro ufficio. Attraverso questo approccio si potrebbe semplificare la gestione della conservazione dei dati, permettendo agli utenti di applicare l'etichetta appropriata anche senza una conoscenza dettagliata delle leggi.

Capitolo 6

Definizione del Modello di Rischio

In questo capitolo, verrà affrontata l'analisi del rischio, svolta come complemento alle attività trattate nel capitolo precedente. Attraverso questo lavoro, sarà possibile valutare e comprendere meglio le minacce e le vulnerabilità che possono influire sulla sicurezza del sistema e il livello di pericolo associato ad esse.

6.1 Metodologia

L'analisi del rischio è stata condotta prendendo spunto dalle linee guida fornite dal NIST [1]. Il documento preso come riferimento è la Special Pubblication (SP) 800-30 intitolata "Guide for Conducting Risk Assessment" [33], facente parte di una serie di pubblicazioni che trattano specificamente l'argomento delle valutazioni del rischio. Questa guida fornisce suggerimenti per condurre valutazioni del rischio in organizzazioni e sistemi di informazione federali, ampliando le indicazioni fornite nella Special Publication 800-39 [15]. Il documento, quindi, fornisce alle organizzazioni un processo per identificare, valutare e mitigare i rischi associati alla sicurezza informatica. La struttura dell'approccio rimane flessibile e adattabile alle specifiche esigenze delle differenti aziende.

La valutazione del rischio, condotta in questo lavoro di tesi, è un'analisi di tipo qualitativo. Dopo un'accurata esplorazione delle potenziali minacce e vulnerabilità del sistema, circoscritte agli argomenti trattati nella tesi, è stata determinata l'entità dell'impatto che

[1] National Institute of Standards and Technology, agenzia governativa degli Stati Uniti che si occupa di promuovere l'innovazione e la competitività industriale attraverso l'istituzione di standard e la diffusione di best practice in vari settori, tra cui la sicurezza informatica.

alcune circostanze o eventi potrebbero avere sull'azienda e la probabilità che questi si verifichino. I fattori di rischio sono stati classificati in uno dei seguenti livelli: molto basso, basso, moderato, alto, molto alto.

Il lavoro è stato sviluppato attraverso diverse fasi:

1. Identificazione delle minacce
2. Identificazione delle vulnerabilità
3. Valutazione della probabilità di occorrenza
4. Valutazione dell'impatto
5. Generazione della matrice di rischio finale

6.1.1 Identificazione delle minacce

Sono stati identificati sia gli eventi che le fonti di minaccia su cui si vuole porre l'attenzione in questa analisi del rischio.

La guida distingue le fonti di minaccia in due principali categorie: "avversarie" e "non avversarie". Il termine Avversario o "Adversary" è da intendere come "individuo, gruppo, organizzazione o governo che conduce o ha l'intenzione di condurre attività dannose" [6]. In questa tesi, l'attenzione è rivolta principalmente all'uso improprio delle risorse informatiche da parte dei dipendenti dell'organizzazione, piuttosto che agli attacchi esterni intenzionalmente dannosi. Pertanto, nello studio del rischio si è scelto di considerare solo le fonti di minaccia non avversarie, con un focus particolare sulle seguenti:

1. Utente standard (Accidentale)
2. Utente con privilegi/Amministratore (Accidentale)

Si prende in considerazione solo il caso in cui entrambe le fonti potrebbero innescare una minaccia in modo accidentale, vale a dire compiendo azioni erronee durante l'esecuzione delle loro responsabilità quotidiane. In Tabella 6.1 viene assegnato a entrambe le fonti un identificativo e un range di effetti, ovvero un'indicazione della varietà di impatti potenziali che una loro azione può avere sul sistema.

Identificativo	Fonte di minaccia	Range di effetti
FM1	Utente standard	Moderato
FM2	Utente con privilegi	Alto

Tabella 6.1. Lista delle fonti di minaccia e range di effetti

Alla fonte FM1 è stato assegnato un valore moderato di range degli effetti, il che indica che un potenziale incidente di sicurezza potrebbe coinvolgere una parte significativa delle risorse, tra cui alcune di tipo critico. D'altra parte, nel caso della fonte FM2, l'incidente avrebbe un impatto più ampio, coinvolgendo un numero maggiore di risorse, tra cui molte di quelle critiche.

Le minacce che sono state prese in considerazione in questo studio del rischio sono state selezionate tra quelle suggerite dalla guida precedentemente citata e sono mostrate in Tabella 6.2, ciascuna con il relativo identificativo. Nell'ultima colonna sono indicate

Identificativo	Minaccia	Fonte
M1	Impostazione dei privilegi errata	FM2
M2	Gestione scorretta di informazioni riservate	FM1/FM2

Tabella 6.2. Lista delle minacce

le fonti che possono concretizzare la minaccia. Un'impostazione errata dei privilegi può essere causata dall'intervento di un utente privilegiato o di un amministratore, che ha i permessi di attribuire o togliere autorizzazioni ad altri utenti per accedere a delle risorse. Un'impropria gestione delle informazioni riservate, invece, può essere attribuita a qualsiasi utente che abbia accesso a tali dati per motivi professionali. Questo può includere sia gli utenti comuni che gli utenti amministratori.

6.1.2 Identificazione delle vulnerabilità

Le vulnerabilità del sistema, elencate in Tabella 6.3, sono state identificate in base al lavoro effettuato sulla classificazione automatica e sul controllo degli accessi. Per ciascuna vulnerabilità viene fornito un identificativo, un livello di gravità e un'indicazione di quali caratteristiche dei dati (tra riservatezza, integrità e disponibilità) potrebbero essere compromesse dalla vulnerabilità.

Le vulnerabilità sono state organizzate in categorie, in base alla loro natura. Le diverse categorie sono indicate da diversi colori: la categoria "Accesso pubblico" è rappresentata in giallo, "Gestione impropria dei gruppi standard" in verde, "Altri tipi di accesso" in arancione e, infine, "Assenza di etichetta" in azzurro.

Le vulnerabilità V1, V2 e V3 sono caratterizzate dalla presenza del gruppo speciale "Everyone" all'interno del gruppo Membri o del gruppo Visitatori. Questo implica che tutti gli utenti che fanno parte del gruppo Everyone, che tipicamente comprende tutti gli utenti interni all'organizzazione, possono accedere al sito con le autorizzazioni da membro o visitatore. Le vulnerabilità che scaturiscono da questa circostanza sono state raggruppate nella categoria "Accesso pubblico". Nei casi in cui esista una ragione valida per l'utilizzo di questo gruppo, come ad esempio l'accesso a siti a cui tutti i dipendenti dell'azienda devono poter entrare, la gravità di questa vulnerabilità è stata considerata

ID	Vulnerabilità	Gravità	Potenziale perdita di RID
V1	Accesso “Everyone” come Membro/Visitatore (giustificazione)	Bassa	Riservatezza, Integrità
V2	Accesso “Everyone” come Membro (no giustificazione)	Alta	Riservatezza, Integrità
V3	Accesso “Everyone” come Visitatore (no giustificazione)	Moderata	Riservatezza
V4	Assenza di Proprietari	Molto Alta	Riservatezza, Disponibilità
V5	Presenza di un solo Proprietario	Alta	Riservatezza, Disponibilità
V6	Elevato numero di Proprietari	Molto Alta	Riservatezza, Integrità
V7	Elevato numero di Membri	Alta	Integrità
V8	Presenza di gruppi personalizzati	Moderata	Riservatezza, Integrità
V9	Utilizzo di numerosi “Shareable Links”	Moderata	Riservatezza
V10	File generali non classificati	Moderata	Riservatezza
V11	File estremamente riservati non classificati	Alta	Riservatezza
V12	File estremamente riservati non classificati	Molto Alta	Riservatezza

Tabella 6.3. Lista delle vulnerabilità. Legenda: “Accesso pubblico”, “Gestione impropria dei gruppi standard”, “Altri tipi di accesso”, “Assenza di etichetta”.

bassa. Se la presenza del gruppo speciale non è giustificata e tale gruppo è inserito in quello dei Visitatori, con permessi di sola lettura, la vulnerabilità assume una gravità moderata, comportando un rischio per la riservatezza dei dati. D'altra parte, se il gruppo Everyone viene inserito nel gruppo Membri, allora tutti gli utenti hanno accesso con i permessi di lettura e modifica. Il verificarsi di tale condizione, senza una valida motivazione, determina una vulnerabilità con gravità alta, poiché chiunque appartenga al gruppo speciale acquisisce delle autorizzazioni che potrebbero portare a compromettere non solo la riservatezza, ma anche l'integrità delle informazioni.

Le vulnerabilità V4, V5, V6, V7 sono racchiuse nella categoria di vulnerabilità relative alla "Gestione impropria dei gruppi standard". In particolare, le prime tre fanno riferimento al numero di Proprietari per sito. Nel primo caso l'assenza totale di Proprietari in un sito implica che nessuno all'interno di quel sito possa più modificarne le impostazioni e aggiungere o rimuovere utenti. Ne potrebbe derivare una perdita della disponibilità delle informazioni se nuovi utenti avessero bisogno di accedere al contenuto del sito, ma non potrebbero essere aggiunti come Membri o Visitatori. Al contrario, se non si fosse in grado di rimuovere utenti che non dovrebbero più avere accesso alle informazioni, si avrebbe una perdita di riservatezza. Per le motivazioni spiegate, la vulnerabilità V4 è stata catalogata come a gravità molto alta. Per quanto riguarda la vulnerabilità V5, pur potendosi fare considerazioni simili al caso precedente, la situazione è leggermente diversa per l'esistenza di un Proprietario del sito. Tuttavia, la gravità è stata valutata come alta per i problemi che potrebbero sorgere nel caso in cui l'unico Proprietario dovesse lasciare il sito. L'ultima vulnerabilità riferita al numero di Proprietari è la V6, a cui è stato assegnato un livello di gravità molto alto per la presenza di un elevato numero di utenti che hanno il controllo totale del sito. Il rischio consiste nella possibilità di un alto numero di utenti autorizzati ad aggiungere Membri nel sito che, se non necessari, potrebbero avere un accesso ingiustificato a contenuti riservati. Per questo motivo, questa vulnerabilità potrebbe compromettere la riservatezza e l'integrità dei dati.

L'ultima vulnerabilità segnalata in verde è la V7, per cui valgono gli stessi concetti della precedente, ma applicati al gruppo dei Membri. Anche in questo caso, la presenza di un elevato numero di membri si traduce nella facoltà di ognuno di operare modifiche al contenuto di un sito, aumentando la probabilità di errori che potrebbero compromettere l'integrità dei dati.

Le vulnerabilità V8 e V9 sono state raggruppate sotto la categoria "Altri tipi di accesso" e sono state classificate entrambe come di gravità moderata. La presenza di gruppi e set di permessi personalizzati, deviando da una gestione standard, può rendere più complesso il controllo degli accessi. Se non giustificata, questa situazione determina una vulnerabilità di gravità moderata. Inoltre, un elevato numero di Shareable Links per un singolo sito aumenta la probabilità che questi collegamenti vengano configurati in modo improprio, permettendo la diffusione dei contenuti del sito a utenti non autorizzati, con la conseguente compromissione della riservatezza dei dati.

L'ultima categoria, indicata in azzurro, è denominata "Assenza di etichette" e include

le vulnerabilità V10, V11 e V12. Tutte e tre sono vulnerabilità correlate alla mancata classificazione manuale da parte degli utenti, che può portare a una perdita di riservatezza delle informazioni. Le vulnerabilità presentano diversi livelli di gravità a seconda del livello di riservatezza associato alle varie etichette.

6.1.3 Valutazione dell'impatto e della probabilità

In questa fase dell'analisi del rischio è stata determinata la probabilità che gli eventi associati alle minacce prese in considerazione possano verificarsi e comportare impatti negativi per l'azienda. Si tratta di un fattore di rischio ponderato basato sull'analisi della probabilità che una determinata minaccia sia in grado di sfruttare una o più vulnerabilità specifiche. Per assegnare un valore qualitativo alla probabilità, sono state prese in considerazione le caratteristiche delle fonti di minaccia che potrebbero innescare gli eventi e le vulnerabilità identificate.

L'attività di valutazione della probabilità complessiva si svolge in tre fasi:

1. Valutazione della probabilità che gli eventi considerati come minacce si verifichino.

2. Valutazione della probabilità che le minacce, una volta avviate, provochino impatti negativi.

3. Combinazione delle probabilità precedenti per ricavare la probabilità complessiva.

Per quanto riguarda l'impatto, sono state identificate le tipologie di impatto che potrebbero essere rilevanti in questa analisi del rischio:

- Danni alle operazioni, ossia qualsiasi impatto negativo che una minaccia potrebbe avere sul normale funzionamento dell'organizzazione. Questi potrebbero includere danni derivanti da non conformità, come sanzioni e multe, o danni alla reputazione.

- Danni alle risorse dell'azienda, come ad esempio la perdita di proprietà intellettuale. Questo può avere un impatto significativo su un'organizzazione, poiché può comportare la perdita di potenziali introiti economici o un vantaggio competitivo nei confronti di altre aziende.

- Danni agli individui, derivanti dalla diffusione di informazioni personali e sensibili.

Con livello d'impatto di una minaccia si intende l'entità del danno che potrebbe derivare dalla divulgazione, modifica, cancellazione non autorizzata di dati o dalla perdita di disponibilità di informazioni. Questo danno potrebbe influire su diverse parti interessate, sia all'interno che all'esterno dell'organizzazione, come, ad esempio, dipendenti, aziende partner e clienti.

6.2 Matrice di Rischio

La determinazione del rischio complessivo è un processo che integra i diversi fattori discussi nella sezione precedente. Questa combinazione produce un livello di rischio complessivo per ciascuna minaccia in relazione alle vulnerabilità ad essa associate. Il risultato finale è rappresentato in una matrice di rischio, uno strumento visivo che facilita l'identificazione e la valutazione dei rischi, tenendo conto della gravità del loro impatto e della probabilità del loro verificarsi.

La matrice di rischio è rappresentata in Figura 6.1, dove la colonna numero 10 rappresenta il livello di rischio complessivo per ciascuna riga della matrice, calcolato in base alla combinazione della probabilità complessiva (colonna 8) con il livello di impatto (colonna 9).

	1	2	3	4	5	6	7	8	9	10
1	MINACCIA	FONTE	RANGE DI EFFETTI	PROBABILITÀ DI OCCORRENZA	VULNERABILITÀ	GRAVITÀ	PROBABILITÀ DI RISULTARE IN IMPATTI AVVERSI	PROBABILITÀ COMPLESSIVA	LIVELLO DI IMPATTO	RISCHIO
2	M1	FM2	Alto	Moderata	V1	Bassa	Molto Bassa	Bassa	Basso	**Basso**
3	M1	FM2	Alto	Moderata	V2	Alta	Alta	Alta	Alto	**Alto**
4	M1	FM2	Alto	Moderata	V3	Moderata	Bassa	Bassa	Moderato	**Basso**
5	M1	FM2	Alto	Moderata	V4	Molto Alta	Molto Alta	Alta	Alto	**Alto**
6	M1	FM2	Alto	Moderata	V5	Alta	Moderata	Moderata	Moderato	**Moderato**
7	M1	FM2	Alto	Moderata	V6	Molto Alta	Moderata	Moderata	Alto	**Moderato**
8	M1	FM2	Alto	Moderata	V7	Alta	Bassa	Bassa	Moderato	**Basso**
9	M1	FM2	Alto	Bassa	V8	Moderata	Moderata	Bassa	Moderato	**Basso**
10	M1	FM2	Alto	Moderata	V9	Moderata	Alta	Moderata	Alto	**Moderato**
11	M2	FM1, FM2	Alto/Moderato	Moderata	V10	Moderata	Bassa	Moderata	Moderato	**Moderato**
12	M2	FM1, FM2	Alto/Moderato	Alta	V11	Alta	Molto Alta	Molto alta	Alto	**Alto**
13	M2	FM1, FM2	Alto/Moderato	Alta	V12	Molto Alta	Molto Alta	Molto alta	Molto Alto	**Molto Alto**

Figura 6.1. Matrice di Rischio finale

Un livello di rischio "Molto Alto" significa che la minaccia potrebbe avere diversi effetti negativi gravi. Un livello "Alto" indica che una minaccia potrebbe avere un effetto negativo grave. "Moderato" vuol dire che la minaccia potrebbe avere un effetto negativo serio per l'organizzazione. Un livello "Basso" significa che una minaccia potrebbe avere un effetto negativo limitato. Infine, livello "Molto Basso" indica che la minaccia potrebbe avere un effetto negativo trascurabile.

Dalla matrice di rischio, si può osservare che, nonostante alcune vulnerabilità siano state classificate come di gravità Alta o Molto Alta, il loro rischio complessivo può risultare di un livello inferiore. Questo perché la valutazione del rischio non si basa unicamente sulla gravità della vulnerabilità, ma prende in considerazione l'intero scenario di minaccia. In altre parole, si tiene conto sia della probabilità che l'evento si verifichi, sia dell'impatto che avrebbe sull'organizzazione se dovesse effettivamente verificarsi.

Ad esempio, si può notare che la minaccia M2 in associazione alla vulnerabilità V12 rappresenta il rischio più alto per l'organizzazione tra i vari casi analizzati. Ciò suggerisce che questa particolare combinazione di minaccia e vulnerabilità dovrebbe essere una priorità per le misure di mitigazione del rischio. Inoltre, le righe 3, 5 e 12 indicano un livello di rischio Alto, pertanto anche queste dovrebbero essere prese in considerazione nelle strategie di gestione del rischio dell'organizzazione.

Nell'ambito di questa analisi, si potrebbe fissare la soglia di tollerabilità del rischio al livello Moderato. Questa soglia rappresenta il massimo livello di rischio che l'azienda è disposta ad accettare. Di conseguenza, le minacce al di sotto di tale soglia potrebbero essere considerate come trascurabili e l'organizzazione potrebbe scegliere di accettare tali rischi piuttosto che investire risorse per mitigarli.

È tuttavia importante sottolineare che il livello di rischio e la sua tollerabilità non sono dei valori statici, ma possono subire variazioni nel tempo in risposta a nuove informazioni, a cambiamenti nell'ambiente operativo o a modifiche agli obiettivi dell'organizzazione. Ne consegue che la valutazione del rischio effettuata in questa fase sarà sottoposta ad una revisione e ad un aggiornamento in base alle informazioni ottenute dai risultati dell'implementazione delle misure di sicurezza sviluppate in questo lavoro.

Capitolo 7

Validazione

In questo capitolo verranno riportati e discussi i risultati dell'applicazione delle strategie implementate, descritte nel capitolo 5. Il livello di dettaglio dei dati riportati varierà a seconda della confidenzialità delle informazioni ricavate. Per una presentazione chiara dei risultati, verrà fatto uso di grafici e tabelle.

Al termine di questo capitolo, verrà presentata una rivalutazione del rischio, calcolato nel capitolo 6, in seguito all'implementazione delle soluzioni progettate e all'analisi dei risultati ottenuti.

7.1 Risultati dell'attivazione delle policy di classificazione Automatica

In questa sezione verranno mostrati e discussi i risultati della classificazione automatica ottenuti sui tre siti SharePoint Online presi come campione per l'attivazione delle policy nella loro configurazione definitiva.

7.1.1 Analisi dei risultati per singoli siti campione

La Tabella 7.1 mostra i risultati dell'analisi delle policy di classificazione automatica applicate al primo sito campione. Per ciascuna estensione di file viene indicato il numero di file totali analizzati, di quelli etichettati manualmente, di quelli etichettati dal client con l'etichetta predefinita, dei file con etichetta predefinita che deve essere aggiornata e di file a cui l'etichetta deve essere applicata per la prima volta. Si evidenzia come solo 15 file su 133, ovvero circa l'11%, sono stati manualmente etichettati. Inoltre, la totalità dei PDF non possiede nè l'etichetta manuale nè quella automatica, per cui risulta da etichettare per la prima volta.

Il grafico a torta nella Figura 7.1 rivela che i 133 file nei formati analizzati dallo strumento di classificazione costituiscono il 27,71% del totale dei documenti presenti. Le

Estensione	File totali	Etichettati manualmente	Etichettati di default	Etichetta da aggiornare	Etichetta da applicare
pdf	9	0	0	0	9
xlsx	62	13	35	1	14
docx	62	2	54	3	6
pptx	0	0	0	0	0
Totale	133	15	89	4	29
%		11,28%	66,92%	3,01%	21,80%

Tabella 7.1. Risultati della classificazione automatica sul primo sito campione

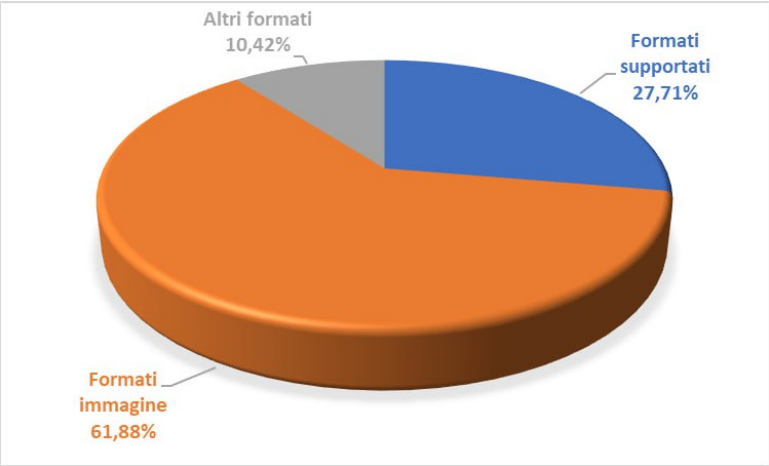

Figura 7.1. Distribuzione dei formati dei file ospitati nel primo sito campione

immagini rappresentano il 61,88% dei file e il 10,42% è costituito da altri formati. Questi dati indicano che la maggior parte dei documenti su questo sito non può essere classificata automaticamente con gli strumenti utilizzati in questo lavoro di tesi.

La Tabella 7.2, analoga alla precedente, mostra i risultati per il secondo sito campione. In questo caso salta subito all'occhio, osservando la terza colonna, che nessun file presente nella libreria di documenti del sito in questione è stato etichettato manualmente. Anche i file etichettati di default sono pochi, solo 5 su 77, quindi il 15,4%.

Nell'ultimo caso, i risultati ottenuti possono essere spiegati dal fatto che numerosi utenti esterni all'organizzazione partecipano al caricamento dei documenti sul secondo sito campione. Questi utenti non hanno la possibilità di applicare manualmente le etichette. Infatti, se i file vengono caricati sul cloud da utenti esterni, l'etichetta predefinita non può essere applicata fino a quando un utente interno non apre il file sul proprio dispositivo. Questo è anche il motivo per cui pochi file risultano etichettati di default.

Estensione	File totali	Etichettati manualmente	Etichettati di default	Etichetta da aggiornare	Etichetta da applicare
pdf	3	0	0	0	3
xlsx	41	0	1	1	42
docx	21	0	4	3	17
pptx	10	0	0	0	10
Totale	77	0	5	4	72
%		0,00%	6,49%	5,19%	93,51%

Tabella 7.2. Risultati della classificazione automatica sul secondo sito campione

Di conseguenza, la classificazione automatica riveste un ruolo particolarmente importante in questo specifico contesto, in quanto permette di etichettare la maggior parte dei file ospitati sul sito che altrimenti rimarrebbero non classificati.

Un altro dato importante riguarda la distribuzione dei formati di file ospitati sul sito, raffigurata in Figura 7.2, da cui si nota una situazione molto diversa rispetto al primo sito campione. In questo caso, su un totale di 15691 file presenti sul sito, solo lo 0,49% può essere analizzato dallo strumento di classificazione automatica adottato.

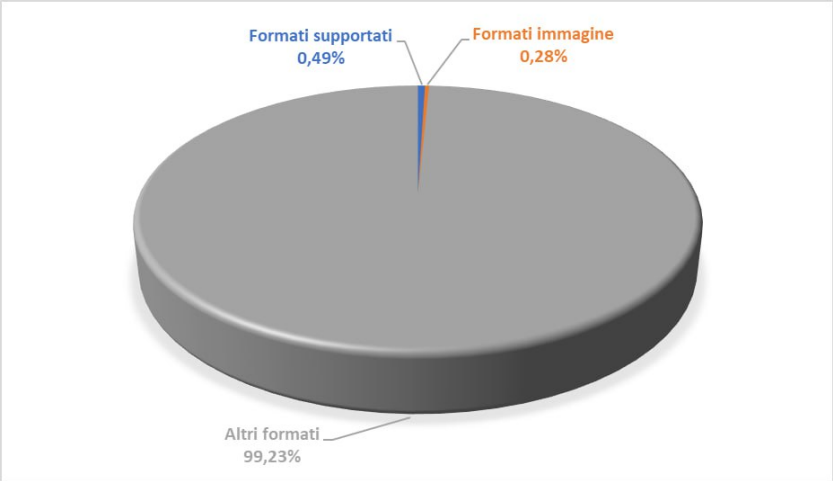

Figura 7.2. Distribuzione dei formati dei file ospitati nel secondo sito campione

In Tabella 7.3, si leggono i dati relativi ai risultati sul terzo sito campione, dove addirittura meno dell'1% dei file è stato etichettato manualmente. Dal grafico in Figura 7.3, si nota nella fetta in blu un numero di file con estensioni supportate molto più alto dei due precedenti in relazione al totale dei documenti ospitati sul sito. In questo caso, l'applicazione della classificazione automatica si rivela particolarmente utile.

Estensione	File totali	Etichettati manualmente	Etichettati di default	Etichetta da aggiornare	Etichetta da applicare
pdf	1967	0	0	0	1967
xlsx	2356	29	379	41	1948
docx	2191	28	474	5	1689
pptx	435	5	46	5	384
Totale	6949	62	899	51	5988
%		0,89%	12,94%	0,73%	86,17%

Tabella 7.3. Risultati della classificazione automatica sul terzo sito campione

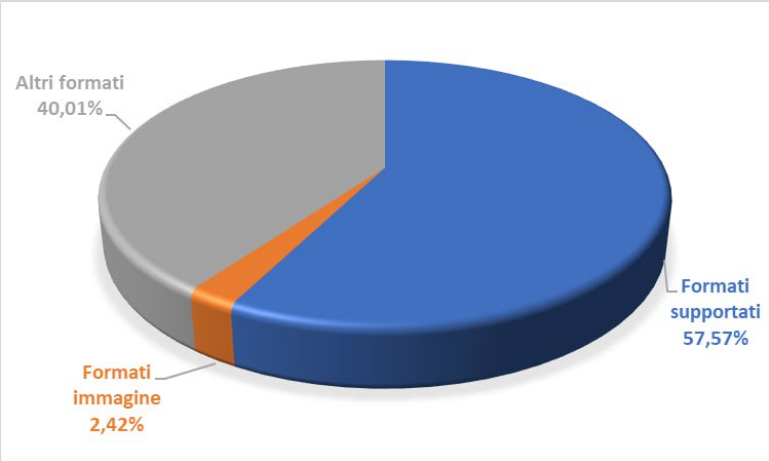

Figura 7.3. Distribuzione dei formati dei file ospitati nel terzo sito campione

7.1.2 Analisi dei risultati complessivi

La variabilità dei risultati ottenuti per ciascun campione evidenzia come la situazione possa differire da sito a sito, a seconda della formazione e della consapevolezza degli utenti che ne fanno parte, della tipologia di documenti che vi sono salvati e in base alla presenza o meno di collaboratori esterni. Questo suggerisce che le statistiche ricavate dalla somma dei risultati dei singoli siti campione potrebbero non riflettere fedelmente la situazione aziendale complessiva.

Il grafico a torta in Figura 7.4 mostra la distribuzione totale dei formati dei file ospitati nei tre siti campione. La porzione colorata di blu rappresenta i formati che lo strumento è stato in grado di analizzare. Nonostante una grande quantità di file nei siti fosse salvata in formati non supportati dalla classificazione, ciò non implica che questi file siano privi di protezione. Infatti, molti di questi file sono protetti attraverso altri strumenti all'interno dell'organizzazione. Un esempio di ulteriore protezione per i file può essere rappresentato dal controllo degli accessi, come dimostrato in questo studio.

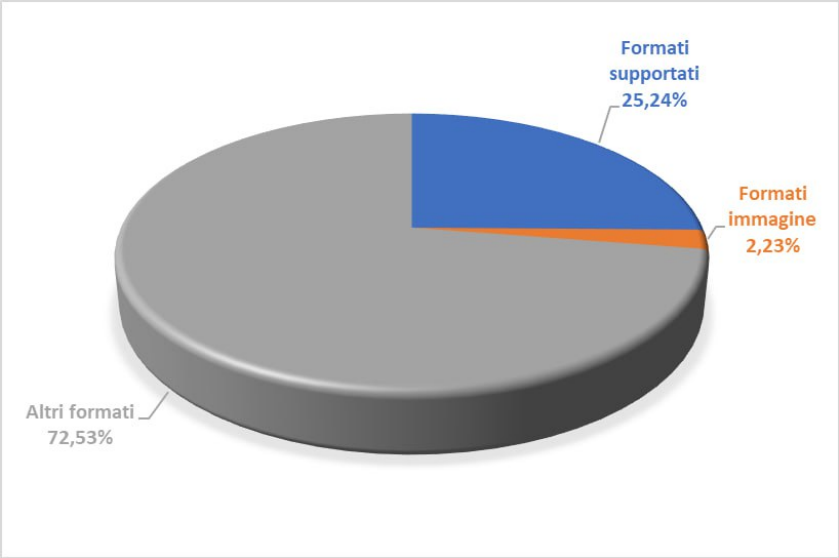

Figura 7.4. Distribuzione complessiva dei formati dei file nel totale dei tre siti campione

In Tabella 7.4, vengono riportati i risultati complessivi della classificazione automatica ottenuti sommando i risultati dei tre siti campione. L'ultima colonna rappresenta la

Estensione	File totali	Etichetta aggiornata	Etichetta applicata per la prima volta	% classificati automaticamente
pdf	1979	0	1979	100%
xlsx	2461	43	2004	83,18%
docx	2274	11	1712	75,77%
pptx	445	5	394	89,66%
Totale	7159	59	6089	85,88%

Tabella 7.4. Risultati complessivi della classificazione automatica

percentuale di file interessati dalla classificazione automatica, sia perchè sia stata applicata un'etichetta con priorità superiore, sia perché sia stata applicata un'etichetta per la prima volta. Quindi il risultato di questa colonna è la somma delle colonne numero 3 e numero 4, considerata come percentuale sul totale dei file analizzati dallo strumento.

L'istogramma in Figura 7.5, raffigura la distribuzione delle etichette che sono state applicate dall'automatismo. Come si può notare, sui siti presi come campione, una piccola quantità di file è stata etichettata come "Estremamente Riservato" (12 file), mentre un numero maggiore è stato classificato come "Riservato" (430 file). La gran parte dei file, al contrario, ha ricevuto l'etichetta "Generale".

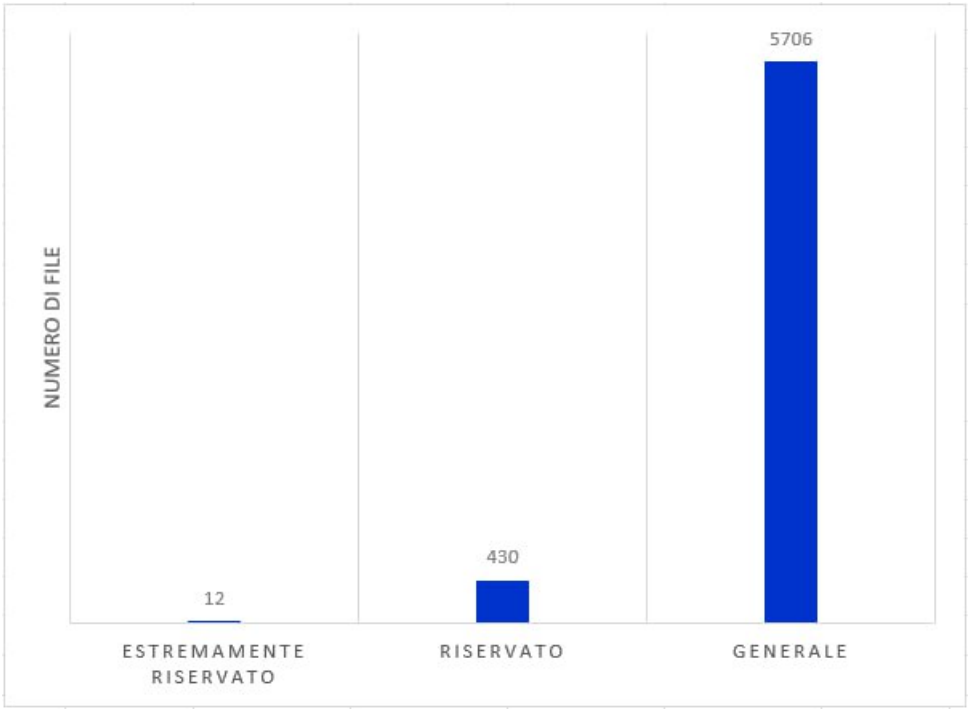

Figura 7.5. Distribuzione delle etichette applicate automaticamente

Dal confronto tra i file etichettati a mano e quelli etichettati automaticamente, mostrato in Figura 7.6, si evidenzia l'efficienza della classificazione automatica contro l'etichettatura manuale. Infatti è importante ricordare che, sebbene la configurazione della classificazione adottata in questo lavoro di tesi non sia in grado di gestire l'etichettatura automatica delle immagini, gli utenti dell'organizzazione hanno la possibilità di farlo manualmente. Nonostante ciò, nessuna delle immagini presenti sul sito è stata classificata manualmente. Un discorso analogo si può rilevare per i formati PDF, che non sono stati in nessun caso etichettati manualmente dagli utenti.

Il fatto che nessuno dei file che gli utenti avrebbero potuto classificare manualmente sia stato effettivamente classificato suggerisce che, in assenza di un processo automatico che gestisca la classificazione o che fornisca suggerimenti su come gestire i documenti, gli utenti tendono non eseguire l'operazione manuale di etichettatura. Questo sottolinea l'importanza di avere un sistema di classificazione automatica per garantire una gestione efficace dei documenti.

Per concludere, anche se il 25% di file analizzati sul totale dei file presenti sui siti potrebbe sembrare una percentuale piccola, in realtà rappresenta 7159 file. Di questi, l'85,88% è stato interessato dalla classificazione automatica, corrispondente a un totale di 6148 file che ora possono essere adeguatamente protetti. Questo numero potrebbe

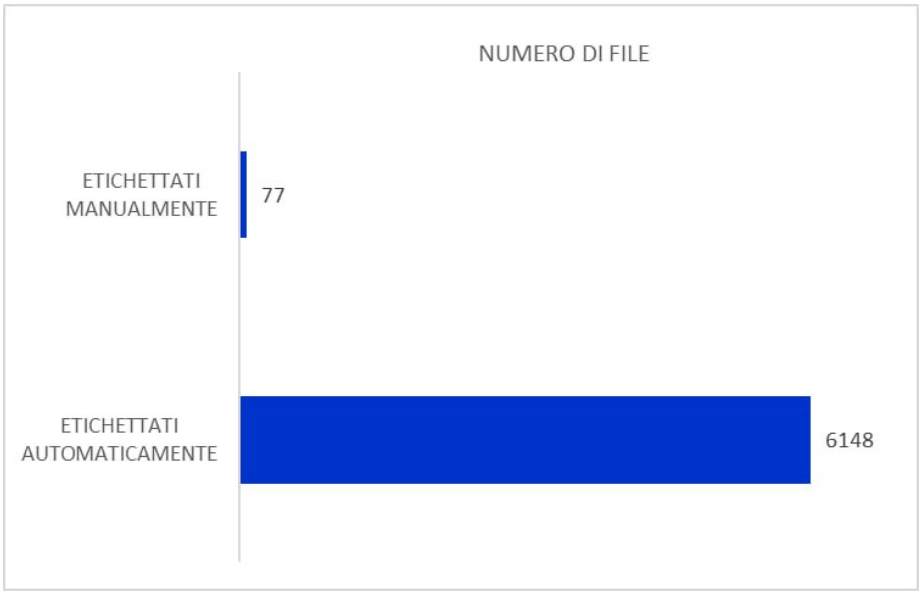

Figura 7.6. Confronto tra il numero di file classificati manualmente e file classificati automaticamente

diventare ancora più significativo con l'estensione delle policy di classificazione automatica a tutti i siti SharePoint dell'organizzazione, con un impatto notevole sulla sicurezza dei dati.

7.2 Risultati del Controllo degli Accessi

In questa sezione verranno mostrati e discussi i risultati ottenuti dall'analisi dei dati raccolti dagli script realizzati per il controllo degli accessi ai siti SharePoint.

Dall'analisi della lista di tutti i siti SharePoint, si è osservata una distribuzione dei modelli di sito all'interno dell'azienda, come viene mostrato nella Tabella 7.5.

Per quanto riguarda i modelli STS#3 e SITEPAGEPUBLISHING#0, i risultati non hanno evidenziato problemi significativi. Infatti, questi siti presentavano caratteristiche

Modello di sito	% di siti sul totale
Siti di team GROUP#0	98,13%
Siti di comunicazione SITEPAGEPUBLISHING#0	0,87%
Siti di team STS#3	0,97%
Altri tipi di siti	0,03%

Tabella 7.5. Distribuzione dei diversi modelli di siti SharePoint aziendali

conformi a un utilizzo pressoché standard. Nei siti con modello STS#3, non è stato riscontrato l'uso di link condivisibili, poiché in questi siti la funzionalità risultava disabilitata. Non sono state rilevate neppure altre tipologie di accesso o gruppi diversi da quelli standard. Anche nei siti con modello SITEPAGEPUBLISHING#0 non sono state rilevate particolari criticità. Molti di questi siti presentavano un numero elevato di visitatori, nell'ordine delle decine di migliaia. Questo comportamento è coerente con la natura del tipo di sito, dove ci si aspetta che un gran numero di utenti sia inserito come visitatore per accedere alle comunicazioni che vengono diffuse tramite il sito.

Poiché i modelli SITEPAGEPUBLISHING#0 e STS#3 costituiscono ciascuno meno dell'1% dei siti presenti in azienda e non hanno mostrato criticità rilevanti, sono stati considerati di minore importanza per l'analisi dettagliata dei risultati. Di conseguenza, l'attenzione si è concentrata sui risultati relativi ai siti del modello GROUP#0.

Di tutti i siti con modello GROUP#0, il 94,81% è un sito SharePoint associato a un team di Teams. Questo implica, come spiegato nel capitolo 5.2, che i gruppi Proprietari e Membri sono popolati con gli utenti Proprietari e Membri del team, mentre il gruppo Visitatori richiede l'aggiunta manuale degli utenti direttamente da SharePoint. Questo potrebbe spiegare perché è stato riscontrato che il 99,44% di siti appartenenti a questo modello sono siti con un conteggio di utenti Visitatori pari a 0.

La Tabella 7.6 offre una visione quantitativa di come si distribuiscono i siti in base al numero di Proprietari. Ogni riga della tabella rappresenta un intervallo specifico del numero di Proprietari. La colonna, invece, mostra la percentuale di siti, rispetto al totale, con un numero di Proprietari che rientra in quel determinato intervallo. A ciascun intervallo è stato associato un livello di rischio, indicato nella colonna centrale della tabella, sulla base di ciò che ci si potrebbe aspettare da un uso appropriato dei permessi di accesso. I numeri che rientrano nei range attesi sono considerati a basso rischio, mentre quelli che superano il range sono associati a livelli di rischio più elevati. Dai dati emerge che nessun sito supera il limite prefissato di 100 Proprietari e che tutti i siti SharePoint aziendali hanno un numero di Proprietari che non supera i 30. Il 45,33% dei siti rientra nella fascia considerata a rischio molto basso, con un numero di Proprietari compreso tra 2 e 10 e una percentuale simile, pari al 47,29%, comprende siti con un solo Proprietario. In quest'ultimo caso, l'abbandono da parte dell'unico Proprietario lascerebbe il sito senza supervisione e nessuno potrebbe più gestirlo. Questo scenario si è già verificato nel 6,17% dei siti, che risultano senza alcun Proprietario, e quindi classificati come ad alto rischio.

La Tabella 7.7 presenta i dati raccolti sulla distribuzione dei siti in base al numero di Membri, in maniera analoga a quanto fatto per i Proprietari. La quasi totalità dei siti ha un numero di Membri inferiore a 50, che è il limite stabilito per considerare il rischio come molto basso.

Riguardo ai dati raccolti sugli altri tipi di accesso ai siti SharePoint, la Tabella 7.8 illustra la distribuzione dei siti in base al numero di utenti trovati in gruppi personalizzati, mentre la Tabella 7.9 presenta i dati raccolti sugli Shareable Links, Limited Access e sul gruppo speciale Everyone. La soglia per un controllo degli accessi gestibile sui gruppi

Numero di Proprietari (P)	Livello di rischio	% di siti sul totale di siti
$P = 0$	Elevato	6,17%
$P = 1$	Alto	47,29%
$1 < P <= 10$	Molto basso	45,33%
$10 < P <= 30$	Basso	1,17%
$30 < P <= 70$	Medio	0,00%
$70 < P <= 100$	Alto	0,00%
$P > 100$	Elevato	0,00%

Tabella 7.6. Distribuzione dei siti per numero di Proprietari nel modello GROUP#0

Numero di Membri (M)	Livello di rischio	% di siti sul totale di siti
$0 <= M < 50$	Molto basso	97,75%
$50 <= M < 100$	Basso	1,43%
$100 <= M < 300$	Medio	0,60%
$300 <= M < 500$	Alto	0,09%
$M >= 500$	Elevato	0,13%

Tabella 7.7. Distribuzione dei siti per numero di Membri nel modello GROUP#0

personalizzati è stata fissata a 50 utenti per ciascun gruppo non standard. Una minima percentuale dei siti, lo 0,04%, ha superato questa soglia, un valore così basso da poter essere ritenuto trascurabile. Come illustrato nella Tabella 7.8, solo i siti che non utilizzano gruppi personalizzati sono stati associati a un livello di rischio molto basso, rappresentando il 99,77% del totale dei siti aziendali. Ciò indica che l'uso di siti personalizzati che si discostano dalle autorizzazioni standard è un evento estremamente raro.

In Tabella 7.9, si nota che le percentuali di utilizzo di Collegamenti Condivisibili (Shareable Links) e di Accessi Limitati (Limited Access) sono molto basse. Nei siti in cui queste funzionalità sono presenti, il numero di collegamenti e Accessi Limitati risulta comunque contenuto. Questo suggerisce che l'utilizzo potrebbe essere motivato da esigenze operative reali, piuttosto che da un uso indiscriminato. Nonostante ciò, sono stati identificati 3 siti

Numero di Altri utenti (A)	Livello di rischio	% di siti sul totale di siti con A	% di siti sul totale di siti
$A = 0$	Molto basso		99,77%
$1 <= A < 10$	Basso	53,00%	0,12%
$10 <= A < 20$	Medio	7,00%	0,02%
$20 <= A < 50$	Alto	22,00%	0,05%
$A >= 50$	Elevato	18,00%	0,04%

Tabella 7.8. Distribuzione dei siti per numero di utenti in gruppi non standard nel modello GROUP#0

Tipo di accesso	% di siti sul totale
Shareable Links	3,03%
Limited Access	4,14%
Gruppo Everyone in Proprietari	0,00%
Gruppo Everyone in Visitatori	0,04%
Gruppo Everyone in Membri	0,68%
Gruppo Everyone in Altri	0,01%

Tabella 7.9. Risultati sulle tipologie di accesso non standard nei siti GROUP#0

in cui sono stati rilevati più di 100 Shareable Links ciascuno. In questi casi, è necessaria un'attenta valutazione: bisogna verificare che l'uso dei collegamenti sia strettamente necessario per le operazioni del sito e, se non più utili, procedere alla loro disattivazione.

L'uso del gruppo speciale "Everyone", è stato riscontrato in percentuali molto basse all'interno dei gruppi standard Visitatori e Membri. La sua presenza in questi gruppi deve essere opportunamente giustificata da esigenze operative aziendali, soprattutto nel caso dei Membri. Nello 0,01% dei siti, il gruppo "Everyone" è stato trovato in gruppi personalizzati, che non corrispondono alle autorizzazioni standard. Nonostante questa percentuale sia minima, è importante non trascurare tale situazione e comprendere quali permessi vengano concessi e perché siano stati assegnati a tutti gli utenti.

Da tutte le osservazioni precedenti emerge che l'implementazione di un metodo automatico per il controllo degli accessi ai siti SharePoint Online aziendali ha fornito una panoramica utile sull'utilizzo di questo strumento da parte dei dipendenti. Nonostante siano state identificate alcune situazioni potenzialmente problematiche, ulteriori indagini sarebbero utili per confermare i rischi rilevati anche sulla base delle motivazioni fornite dai Proprietari sulle anomalie individuate nella gestione degli accessi ai loro siti. Le risposte

ottenute darebbero una visione più chiara sui dati da ricercare nei controlli riguardanti le situazioni a rischio e aiuterebbero a capire quali situazioni, apparentemente a rischio, possono essere trascurate perché innocue. Tutto ciò allo scopo di ottimizzare la raccolta delle informazioni utili e affinare i controlli futuri.

7.3 Stima del Rischio

In questa sezione, viene presentata una revisione del rischio analizzato nel capitolo 6. Questa revisione ha tenuto conto dei risultati ottenuti dall'implementazione dei controlli di sicurezza proposti nel corso del lavoro di tesi.

La Figura 7.7 mostra la matrice di rischio aggiornata. Le righe evidenziate in blu sono quelle interessate dalla modifica sulla base dell'analisi dei risultati. La colonna modificata

	1	2	3	4	5	6	7	8	9	10
1	MINACCIA	FONTE	RANGE DI EFFETTI	PROBABILITÀ DI OCCORRENZA	VULNERABILITÀ	GRAVITÀ	PROBABILITÀ DI RISULTARE IN IMPATTI AVVERSI	PROBABILITÀ COMPLESSIVA	LIVELLO DI IMPATTO	RISCHIO
2	M1	FM2	Alto	Bassa	V1	Bassa	Molto Bassa	Molto Bassa	Basso	**Molto Basso**
3	M1	FM2	Alto	Bassa	V2	Alta	Alta	Moderata	Alto	**Moderato**
4	M1	FM2	Alto	Bassa	V3	Moderata	Bassa	Bassa	Moderato	Basso
5	M1	FM2	Alto	Moderata	V4	Molto Alta	Molto Alta	Alta	Alto	Alto
6	M1	FM2	Alto	Moderata	V5	Alta	Moderata	Moderata	Moderato	Moderato
7	M1	FM2	Alto	Bassa	V6	Molto Alta	Moderata	Bassa	Alto	**Basso**
8	M1	FM2	Alto	Moderata	V7	Alta	Bassa	Bassa	Moderato	Basso
9	M1	FM2	Alto	Bassa	V8	Moderata	Moderata	Bassa	Moderato	Basso
10	M1	FM2	Alto	Bassa	V9	Moderata	Alta	Moderata	Alto	Moderato
11	M2	FM1, FM2	Alto/Moderato	Bassa	V10	Moderata	Bassa	Bassa	Moderato	**Basso**
12	M2	FM1, FM2	Alto/Moderato	Moderata	V11	Alta	Molto Alta	Alta	Alto	Alto
13	M2	FM1, FM2	Alto/Moderato	Moderata	V12	Molto Alta	Molto Alta	Alta	Molto Alto	Molto Alto

Figura 7.7. Stima del Rischio dopo l'analisi dei risultati

sulla base delle nuove conoscenze e considerazioni è la 4, colorata in viola e relativa alla "Probabilità di Occorrenza". I cambiamenti in questa colonna, analizzati riga per riga, sono i seguenti:

Righe 2, 3, 4 : La probabilità di occorrenza è stata cambiata da Moderata a Bassa. Questo perché la presenza del gruppo "Everyone" è risultata in percentuali bassissime.

Riga 7 : La probabilità di occorrenza è stata cambiata da Moderata a Bassa perché i risultati hanno evidenziato l'assenza di siti con un numero di Proprietari oltre la soglia stabilita.

Riga 10 : La probabilità di occorrenza è stata ridimensionata da Moderata a Bassa perché è stata trovata una percentuale molto bassa di siti che fanno uso di Shareable Links.

Riga 11 : La probabilità di occorrenza è stata cambiata da Moderata a Bassa. Questo perché si presume che, con l'adozione della classificazione automatica, il numero di file generali non classificati sia limitato e ristretto esclusivamente ai formati di file non supportati dall'automazione.

Righe 12, 13 : La probabilità di occorrenza è stata ridimensionata da Alta a Moderata. Questo perché, con l'introduzione della classificazione automatica, si prevede che una considerevole quantità di file nei formati supportati venga correttamente classificata come "Riservato" ed "Estremamente Riservato". La presenza di formati non supportati e di informazioni sensibili che il classificatore non è in grado di identificare non consentono di classificare come bassa questa probabilità.

Anche l'altra colonna colorata in viola, la numero 8, che rappresenta la "Probabilità complessiva", ha subito delle modifiche, poiché è derivata dalla combinazione della colonna 4 e della colonna 7.

Il rischio, indicato nell'ultima colonna, è stato quindi ricalcolato combinando la colonna 8 con la colonna 9, ottenendo così i nuovi risultati. Le celle evidenziate in grassetto nell'ultima colonna sono quelle che hanno subito modifiche, corrispondenti alle righe 2, 3, 7 e 11. In tutti questi casi, si osserva una stima di rischio inferiore rispetto a quella prevista nella prima analisi.

7.4 Sviluppi futuri

Questa sezione presenta alcune proposte per un eventuale sviluppo e ampliamento del progetto trattato in questa tesi.

Un obiettivo prioritario sarebbe quello di completare l'implementazione della conservazione automatica dei dati, monitorando le sue prestazioni e analizzando il beneficio per l'utente.

Riguardo alla classificazione automatica dei dati, si potrebbe integrare il lavoro già svolto con l'utilizzo di classificatori personalizzati, sempre basati su espressioni regolari, che riconoscano qualche tipologia aggiuntiva di dato che segue un pattern ben definito e che per l'azienda sia da proteggere. Tale intervento sarebbe fondamentale per identificare e proteggere efficacemente le informazioni sensibili aggiuntive rispetto a quelle già rilevate. Inoltre, si potrebbe valutare l'utilizzo del Machine Learning per l'individuazione di

specifiche tipologie di documenti comuni in azienda. L'idea è quella di addestrare modelli su documenti appositamente creati, contenenti informazioni significative per l'azienda ma privi di dati sensibili reali. Ciò potrebbe migliorare notevolmente la copertura di file che si riescono a classificare correttamente. Inoltre, potrebbe risultare efficace l'implementazione di un sistema automatico di notifiche in seguito alla classificazione dei documenti a più alto livello di riservatezza. Gli utenti verrebbero avvisati quando un documento da loro creato o modificato viene classificato in modo automatico come estremamente riservato. Questo consentirebbe all'utente di valutare immediatamente se è necessaria l'applicazione della crittografia al file per restringere l'accesso dei contenuti solo a specifiche persone.

Per rafforzare ulteriormente la sicurezza, si potrebbero prevedere revisioni periodiche del rischio integrando all'analisi qualitativa anche quella quantitativa. Questo fornirebbe una visione più ampia e oggettiva del livello di rischio associato ai dati aziendali anche in termini di perdita economica.

Un'ulteriore prospettiva potrebbe riguardare l'associazione di specifici livelli di rischio a ciascun sito SharePoint attraverso l'integrazione delle informazioni derivate dall'applicazione automatica delle etichette di classificazione ai file contenuti in essi, insieme ai dati ottenuti dal monitoraggio degli accessi. Tale approccio consentirebbe di attribuire un grado di riservatezza e, di conseguenza, di rischio, a ciascun sito SharePoint, basandosi sui tipici contenuti memorizzati e condivisi all'interno di essi. Questa pratica semplificherebbe notevolmente la gestione complessiva dei siti, offrendo una visione più chiara e dettagliata della loro rilevanza e criticità all'interno del contesto aziendale.

Infine, si può pensare all'implementazione di strategie di eliminazione automatica dei file che non sono soggetti alle normative sulla conservazione. Si può valutare la definizione di apposite etichette di cancellazione automatica per quei dati che i dipendenti salvano in repository cloud personali e che, se dimenticati, potrebbero essere conservati inutilmente. Questo consentirebbe di distinguere i documenti personali da quelli lavorativi, evitando anche sprechi di spazio di archiviazione.

Capitolo 8

Conclusioni

Nella tesi è stato affrontato il tema dell'automatizzazione dei controlli di sicurezza, con l'obiettivo di migliorare la protezione delle informazioni all'interno dell'azienda oggetto dello studio. Le soluzioni proposte rappresentano un passo avanti verso questo obiettivo, delineando un approccio di sicurezza che accompagna i dati lungo diverse fasi del loro ciclo di vita.

L'attenzione è stata dedicata ai dati ospitati nel cloud, dato il crescente utilizzo di questa tecnologia da parte delle aziende. Grazie alle enormi capacità di archiviazione che le piattaforme cloud mettono a disposizione, si assiste ad una tendenza ad accumulare sempre più informazioni di vario genere. La proliferazione incontrollata dei dati ha sottolineato l'urgenza di un'attenzione e di una gestione più efficace degli stessi, motivando l'intervento su più fronti affrontato in questo lavoro.

Nella tesi è stato dimostrato come la classificazione automatica delle informazioni riservate, al momento della creazione e memorizzazione nel cloud, rappresenta uno strumento utile di supporto per gli utenti, specialmente considerando la comune tendenza a trascurare l'operazione manuale. Inoltre, è emerso che il monitoraggio delle autorizzazioni concesse agli utenti per accedere alla piattaforma cloud può rivelare dati utili sul comportamento degli utenti nella configurazione dei permessi di accesso ai dati. Sono state poste le basi per la progettazione di una strategia di conservazione dei dati in conformità con i periodi previsti dalla legge. L'analisi condotta sul rischio informatico del sistema, limitatamente agli aspetti discussi nel lavoro di tesi, ha contribuito ad arricchire la comprensione delle problematiche di sicurezza trattate e può essere utilizzata per ristabilire la priorità dei controlli sulla base dei nuovi dati acquisiti.

Le soluzioni proposte hanno prodotto una serie di risultati che possono essere ulteriormente esplorati per perfezionare l'utilizzo degli strumenti impiegati e per ottimizzare i controlli implementati, adattandoli sempre più alle specifiche esigenze dell'azienda.

Questo lavoro rappresenta un tentativo di miglioramento della protezione dei dati e può essere ulteriormente sviluppato. Lo studio può estendersi in futuro per coprire aspetti

aggiuntivi della sicurezza attualmente trascurati in molte realtà aziendali, come la gestione automatica dell'eliminazione dei file al termine del periodo di conservazione.

L'adozione di tecniche innovative e sempre più automatizzate può agevolare la protezione delle informazioni, anche se, soprattutto nelle situazioni più critiche, rimane fondamentale che gli utenti mantengano ruolo attivo nella gestione dei propri dati.

In conclusione, la tesi sottolinea l'importanza dell'integrazione di automatismi nella sicurezza dei dati, riconoscendo al contempo la necessità di considerare le limitazioni delle soluzioni proposte, che non possono completamente sostituire il controllo umano.

Bibliografia

[1] Decreto legislativo 21 novembre 2007, n. 231. Pubblicato in "Gazzetta Ufficiale", n. 290 del 14 dicembre 2007 - Supplemento ordinario. URL https://www.gazzettaufficiale.it/eli/id/2007/12/14/007X0246/sg.

[2] Shafi Ahmad, Dillidorai Arumugam, Srdan Bozovic, Elnata Degefa, Sailesh Duvvuri, Steven Gott, Nitish Gupta, Joachim Hammer, Nivedita Kaluskar, Raghav Kaushik, Rakesh Khanduja, Prasad Mujumdar, Gaurav Malhotra, Pankaj Naik, Nikolas Ogg, Krishna Kumar Parthasarthy, Raghu Ramakrishnan, Vlad Rodriguez, Rahul Sharma, Jakub Szymaszek, and Andreas Wolter. Microsoft purview: A system for central governance of data. *Proc. VLDB Endow.*, 16(12):3624–3635, aug 2023. ISSN 2150-8097. doi: 10.14778/3611540.3611552. URL https://doi.org/10.14778/3611540.3611552.

[3] Amazon Web Services (AWS). Cos'è il cloud computing? URL https://aws.amazon.com/it/what-is-cloud-computing/.

[4] Long Cheng, Fang Liu, and Danfeng (Daphne) Yao. Enterprise data breach: causes, challenges, prevention, and future directions. *WIREs Data Mining and Knowledge Discovery*, 7(5):e1211, 2017. doi: https://doi.org/10.1002/widm.1211. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1211.

[5] CISA. What is cybersecurity? URL https://www.cisa.gov/news-events/news/what-cybersecurity.

[6] DHS Risk Steering Committee. Dhs risk lexicon. https://www.cisa.gov/resources-tools/resources/dhs-risk-lexicon, December 2020.

[7] Attorney-General's Department. Privacy act review report. URL https://www.ag.gov.au/sites/default/files/2023-02/privacy-act-review-report_0.pdf.

[8] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02016R0679-20160504.

[9] International Organization for Standardization. Iso/iec 27018:2020. Technical report, 2020. URL https://bsol-bsigroup-com.ezproxy.biblio.polito.it/Bibliographic/BibliographicInfoData/000000000030405001.

[10] J. Gemson Andrew Ebenezer and S. Durga. Big data analytics in healthcare: A survey. *ARPN Journal of Engineering and Applied Sciences*, 10(8):3645 – 3650, 2015. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-84929347119&partnerID=40&md5=d4b610396e6427b12b805526fa89d854. Cited by: 16.

[11] Google. Labels overview. URL https://developers.google.com/drive/api/guides/about-labels.

[12] Object Management Group. *Business Process Model and Notation*. URL https://www.omg.org/spec/BPMN.

[13] IBM. Cos'è il cloud computing? URL https://www.ibm.com/it-it/topics/cloud-computing.

[14] IBM. Cost of a data breach report 2023. Technical report, IBM, 2023. URL https://www.ibm.com/reports/data-breach.

[15] Joint Task Force Transformation Initiative. Managing information security risk: Organization, mission, and information system view. Technical Report NIST Special Publication (SP) 800-39, National Institute of Standards and Technology, Gaithersburg, MD, March 2011.

[16] International Organization for Standardization. ISO 31000:2018(en), Risk management — Guidelines, 2018. URL https://www.iso.org/obp/ui/#iso:std:iso:31000:ed-2:v1:en. [Online; accessed 28-November-2023].

[17] Jahoon Koo, Giluk Kang, and Young-Gab Kim. Security and privacy in big data life cycle: A survey and open challenges. *Sustainability*, 12(24), 2020. ISSN 2071-1050. doi: 10.3390/su122410571. URL https://www.mdpi.com/2071-1050/12/24/10571.

[18] Mike Loukides. *What is data science?* " O'Reilly Media, Inc.", 2011.

[19] Microsoft. Azure active directory, . URL https://learn.microsoft.com/it-it/dotnet/architecture/cloud-native/azure-active-directory.

[20] Microsoft. Informazioni sulle etichette di riservatezza, . URL https://learn.microsoft.com/it-it/purview/sensitivity-labels#what-a-sensitivity-label-is.

[21] Microsoft. Microsoft purview information protection, . URL https://learn.microsoft.com/it-it/purview/information-protection.

[22] Microsoft. Che cos'è powershell?, . URL https://learn.microsoft.com/it-it/powershell/scripting/overview?view=powershell-7.3.

[23] Microsoft. Informazioni sui criteri e sulle etichette di conservazione, . URL https://learn.microsoft.com/it-it/purview/retention?tabs=table-overriden.

[24] Microsoft. Ulteriori informazioni sui tipi di informazioni riservate, . URL https://learn.microsoft.com/it-it/purview/sensitive-information-type-learn-about.

[25] Microsoft. Definizioni delle entità tipo di informazioni sensibili, . URL https://learn.microsoft.com/it-it/purview/sensitive-information-type-entity-definitions.

[26] Microsoft. Sharepoint, . URL https://www.microsoft.com/it-it/microsoft-365/sharepoint/collaboration.

[27] Microsoft. Panoramica dell'integrazione di teams e sharepoint, . URL https://learn.microsoft.com/it-it/sharepoint/teams-connected-sites.

[28] State of California Department of Justice. California consumer privacy act (ccpa). URL https://www.oag.ca.gov/privacy/ccpa.

[29] National Institute of Standards and Technology. Fips 200, minimum security requirements for federal information and information systems. Technical report, National Institute of Standards and Technology, March 2006. URL https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.200.pdf.

[30] Office of the Privacy Commissioner of Canada. Pipeda in brief. URL https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/pipeda_brief/.

[31] Garante per la Protezione dei Dati Personali. Data breach - violazioni di dati personali. URL https://www.garanteprivacy.it/data-breach.

[32] Agenzia per l'Italia Digitale (AGID). Linee guida sulla formazione, gestione e conservazione dei documenti informatici. URL https://www.agid.gov.it/sites/default/files/repository_files/linee_guida_sul_documento_informatico.pdf.

[33] Ronald Ross. Guide for conducting risk assessments. Technical Report NIST Special Publication (SP) 800-30, Rev. 1, National Institute of Standards and Technology, Gaithersburg, MD, September 2012.

[34] Verizon. 2023 data breach investigations report (dbir). Technical report, 2023. URL https://www.verizon.com/business/resources/T377/reports/2023-data-breach-investigations-report-dbir.pdf.

[35] Xiaojun Yu and Qiaoyan Wen. A view about cloud data security from data life cycle. In *2010 International Conference on Computational Intelligence and Software Engineering*, pages 1–4, 2010. doi: 10.1109/CISE.2010.5676895.

[36] Hongjun Zhang, Shuyan Cheng, Qingyuan Cai, and Xiao Jiang. Privacy security protection based on data life cycle. In *2022 World Automation Congress (WAC)*, pages 433–436, 2022. doi: 10.23919/WAC55640.2022.9934483.