# Zen Class

AWS Architecture Overview

## Introduzione

### Scopo del documento

Il presente documento fornisce un'analisi tecnica dettagliata dell'architettura e delle funzionalità della piattaforma ZenClass, un sistema di e-learning basato su AWS con un'integrazione avanzata di intelligenza artificiale (NLP, LLM, RAG).

L'obiettivo è descrivere i principali componenti infrastrutturali, le strategie di sicurezza adottate e le best practice per garantire conformità alle normative e scalabilità del sistema. In particolare, il documento si focalizza su:

- L'architettura della piattaforma su AWS, con un approfondimento sui servizi utilizzati per la gestione dei contenuti, l'autenticazione e l'elaborazione dei dati.
- Le funzionalità di intelligenza artificiale implementate per la trascrizione automatica, la generazione di contenuti e l'interazione tramite chatbot.
- Le misure di sicurezza e conformità adottate, con particolare riferimento alle normative GDPR e agli standard ISO 27001.

### Ambito piattaforma

ZenClass è progettata per offrire un sistema di gestione dell'apprendimento altamente scalabile e sicuro. La piattaforma è destinata a professori e studenti, consentendo l'accesso a contenuti didattici interattivi e supportando l'apprendimento tramite strumenti avanzati di AI.

Funzionalità per i professori

- Accesso autenticato tramite Active Directory.
- Caricamento di lezioni in formato video e testuale.
- Automazione del processo di elaborazione dei contenuti, comprendente:
    - Trascrizione automatica dei video in testo.
    - Estrazione di testo da documenti.
    - Generazione di contenuti basata su modelli LLM conformi al GDPR.
    - Transcodifica dei video per uno streaming.
    - Archiviazione ottimizzata per la gestione dei contenuti grezzi e processati.

Funzionalità per studenti:

- Accesso autenticato tramite Active Directory, garantendo un controllo centralizzato sugli accessi.
- Possibilità di consultare materiali didattici sotto forma di video, documenti e contenuti interattivi.
- Interazione con un chatbot basato su modelli di intelligenza artificiale, in grado di fornire approfondimenti personalizzati sui contenuti dei corsi.

Funzionalità del chatbot:

- Recupero e indicizzazione dei contenuti attraverso un database vettoriale.
- Generazione di risposte tramite LLM conformi agli standard di privacy e sicurezza.

### Obiettivi di Sicurezza

L'infrastruttura della piattaforma è progettata per garantire la massima sicurezza, in conformità con le normative vigenti in materia di protezione dei dati. Le principali misure adottate riguardano:

- Gestione delle Identità e degli Accessi (IAM)
    - Integrazione con Active Directory per la gestione centralizzata degli utenti.
    - IAM con policy granulari per limitare l'accesso ai dati e ai servizi cloud.
- Protezione della Rete
    - Segmentazione dell'infrastruttura su VPC dedicate, con separazione tra frontend, backend e sistemi AI.
    - AWS WAF e AWS Shield per la protezione da attacchi DDoS e minacce web.
    - Principio del "Zero Trust Security Model", con accessi limitati al minimo necessario per ogni componente.
- Protezione dei Dati e Conformità GDPR
    - Crittografia dei dati a riposo e in transito mediante AWS KMS e TLS 1.3.
    - Implementazione di policy di data retention per la gestione dei dati sensibili.
    - Monitoraggio e audit logging tramite AWS CloudTrail e AWS Security Hub, garantendo la tracciabilità degli accessi e delle operazioni sui dati.
- AI Governance e Privacy
    - Mitigazione del rischio di bias nei modelli AI mediante processi di auditing e testing controllato.
    - Utilizzo di LLM con funzionalità di Explainability, per garantire trasparenza nelle risposte fornite dal chatbot.
    - Adozione di meccanismi di anonimizzazione e pseudonimizzazione, per evitare il trattamento non autorizzato di dati personali.

## Tecnologie utilizzate

L'infrastruttura di ZenClass è interamente basata su AWS, con l'impiego di servizi cloud avanzati per la gestione di dati, intelligenza artificiale e sicurezza.

Gestione dei contenuti e streaming:

- Amazon S3, per l'archiviazione scalabile di contenuti didattici.
- AWS Elemental MediaConvert per la transcodifica e ottimizzazione dello streaming.
- Amazon CloudFront per la distribuzione globale dei contenuti video.

AI e NLP:

- Amazon Textract per l'estrazione di testo dai documenti.
- Amazon Bedrock / SageMaker per la generazione di contenuti basati su modelli LLM.
- Amazon OpenSearch per l'indicizzazione dei contenuti e la ricerca semantica basata su RAG.

Autenticazione e Sicurezza:

- AWS IAM Identity Center con Active Directory per la gestione degli accessi.
- AWS WAF e AWS Shield per la protezione dagli attacchi informatici.
- AWS KMS e Secret Manager per la crittografia dei dati e la conservazione sicura delle chiavi.

Backend e computazione scalabile:

- Amazon Lambda per l'orchestrazione dei flussi di lavoro AI/NLP.
- Amazon API Gateway per la gestione sicura delle API della piattaforma.
- Amazon DynamoDB e RDS per la gestione dei dati strutturati.

# Architettura della piattaforma

## Panoramica generale

La piattaforma ZenClass è implementata su AWS seguendo un'architettura scalabile, resiliente e sicura. L'infrastruttura è progettata per supportare la gestione dei contenuti didattici, la distribuzione in streaming e l'integrazione con moduli di intelligenza artificiale per l'indicizzazione semantica dei documenti e il Q&A tramite chatbot.

L'architettura è suddivisa in più livelli funzionali per garantire modularità e sicurezza:

- Livello di autenticazione: gestione degli accessi e dell'identità utente.
- Livello applicativo: gestione del portale.
- Livello AI: integrazione con moduli di intelligenza artificiale per l'analisi e la generazione di contenuti.
- Livello media: ottimizzazione della fruizione dei contenuti tramite una rete di distribuzione globale.

## Livello di autenticazione

L'integrazione tra ADFS, AWS Identity Center e AWS Cognito segue un flusso di autenticazione basato su SAML 2.0 e JWT, garantendo un accesso federato sicuro alla piattaforma ZenClass. Questo meccanismo consente a professori, studenti e amministratori di autenticarsi utilizzando le loro credenziali Active Directory, ottenendo successivamente un JWT da AWS Cognito per accedere alle risorse della piattaforma.


Inizio processo di autenticazione:

- L'utente accede alla piattaforma ZenClass tramite l'interfaccia web o mobile.
- L'applicazione frontend reindirizza l'utente verso AWS Cognito, che funge da Identity Broker.
- AWS Cognito riconosce che l'utente è un membro di Active Directory e reindirizza la richiesta a AWS Identity Center per il processo di autenticazione federata.

Richiesta autenticazione a Identity center:

- AWS Identity Center è configurato per federare gli utenti tramite ADFS.
- AWS Identity Center reindirizza l'utente verso la pagina di login di ADFS, inviando una richiesta di autenticazione SAML.

Autenticazione tramite ADFS:

- L'utente inserisce le credenziali aziendali (username e password di Active Directory).
- ADFS interroga Active Directory (AD) per verificare l'identità dell'utente.
- Se l'autenticazione è valida, ADFS genera un Assertion SAML e lo invia a AWS Identity Center.
    - AWS Identity Center riceve il token SAML da ADFS.
    - Il token viene verificato e validato.
    - Se il token è valido, AWS Identity Center reindirizza l'utente a AWS Cognito, passando il token SAML.

Conversione del token SAML in JWT:

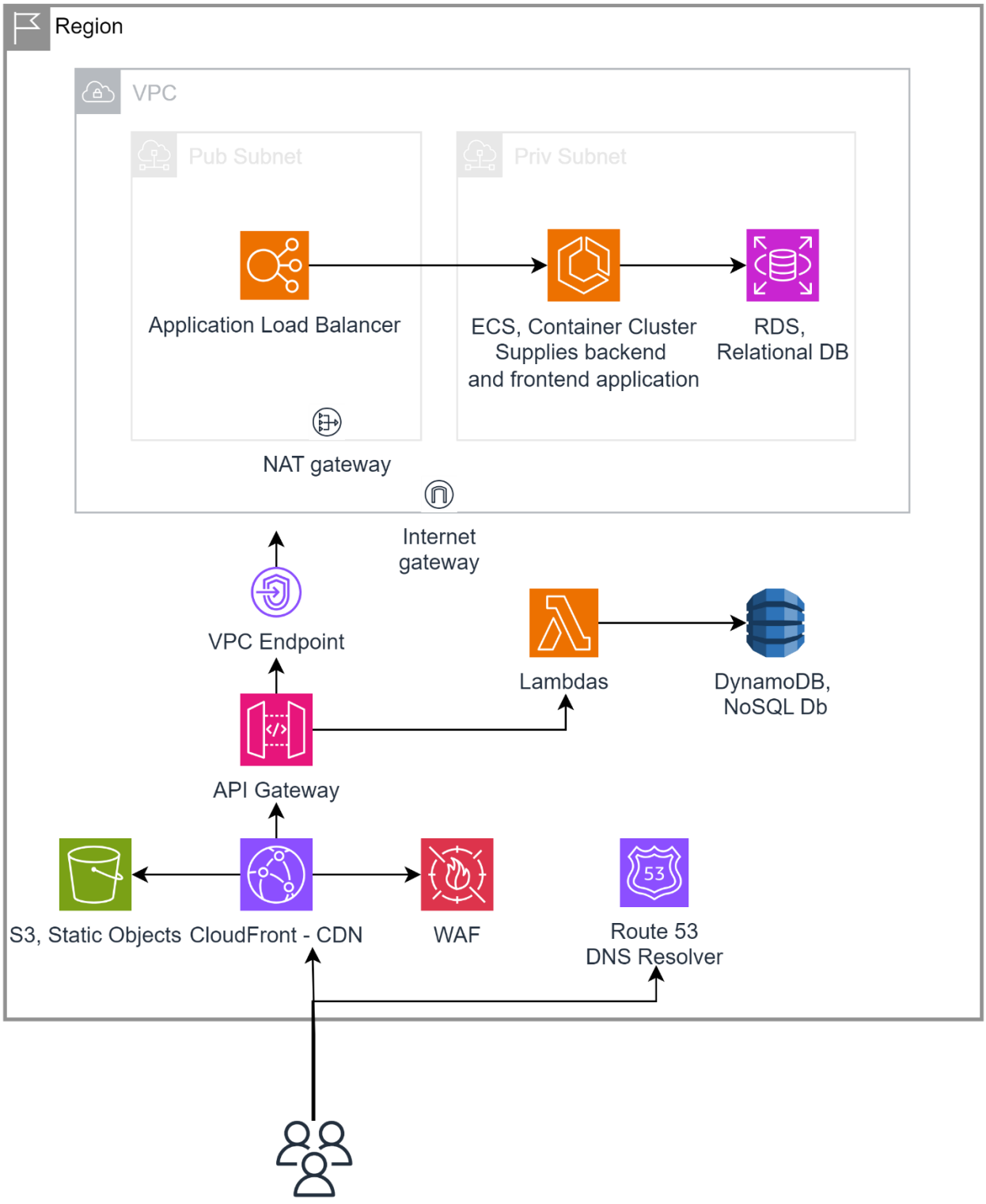
- AWS Cognito riceve il token SAML e lo converte in un JWT (JSON Web Token).

- Cognito genera tre token:
    - ID Token (JWT) → Contiene informazioni sull'utente (nome, email, ruoli).
    - Access Token (JWT) → Utilizzato per autenticare l'utente nelle API della piattaforma ZenClass.
    - Refresh Token → Permette di ottenere nuovi ID Token e Access Token senza dover ripetere il login.

Accesso ai servizi della piattaforma:

- L'applicazione frontend riceve il JWT da AWS Cognito e lo utilizza per effettuare chiamate alle API della piattaforma.
- Ogni richiesta al backend è accompagnata dall'Access Token (JWT), che viene verificato da Amazon API Gateway o direttamente dai servizi backend.
- Il backend valida il token tramite AWS Cognito User Pool prima di concedere l'accesso ai dati.

## Livello applicativo




L'architettura applicativa della piattaforma ZenClass è progettata per gestire richieste utente in modo efficiente, distribuire il carico di lavoro e garantire sicurezza e scalabilità. Il flusso delle richieste si articola attraverso diversi servizi AWS, che interagiscono tra loro per offrire una risposta rapida e affidabile.

Quando un utente accede alla piattaforma, il traffico viene gestito dall'Application Load Balancer (ALB), che smista le richieste in base alla loro natura. Le pagine statiche vengono servite direttamente da Amazon S3 e

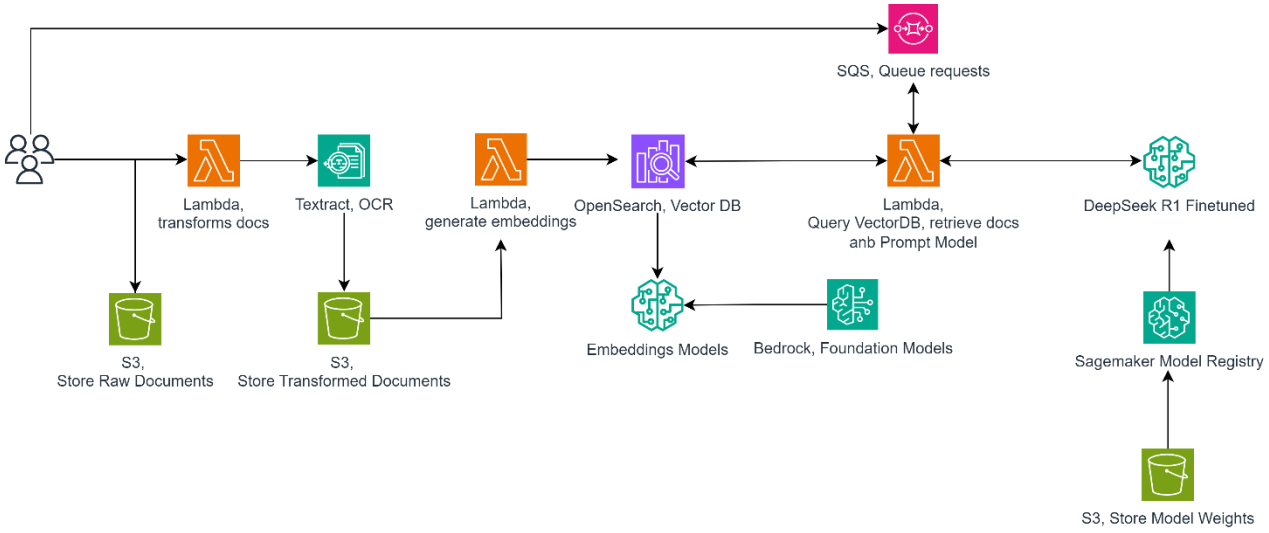
CloudFront, mentre le richieste dinamiche vengono instradate verso Amazon API Gateway o direttamente a ECS, dove il backend elabora i dati.

L'autenticazione è gestita da AWS Cognito, che convalida i token JWT e permette agli utenti autenticati di interagire con le API della piattaforma. API Gateway si occupa di instradare le richieste ai servizi appropriati, proteggendole con rate limiting e autenticazione, mentre il backend containerizzato su ECS esegue la logica applicativa, gestendo operazioni sui database o attivando funzioni AWS Lambda per elaborazioni asincrone (come chiamate a modelli di AI).

I dati vengono recuperati da Amazon RDS per informazioni strutturate e da DynamoDB per richieste ad alta velocità (dati non strutturati, metadati). L'accesso ai database avviene attraverso connessioni sicure all'interno della VPC privata, senza esporre direttamente le risorse su Internet.

Sul fronte della sicurezza, l'architettura prevede protezioni a più livelli. AWS WAF blocca attacchi web, mentre i Security Group e le ACL regolano il traffico tra i servizi. Le API sono protette con autenticazione Cognito e API Gateway applica limiti di traffico per prevenire abusi. Il monitoraggio è garantito da CloudWatch, CloudTrail assicurando il rilevamento di anomalie e la tracciabilità degli accessi.

## Livello AI




L'architettura AI della piattaforma ZenClass è progettata per gestire l'elaborazione documentale, l'indicizzazione vettoriale e l'inferenza tramite modelli di linguaggio. Il flusso dei dati segue un processo strutturato che trasforma documenti grezzi in rappresentazioni ottimizzate per il retrieval e l'interrogazione tramite modelli LLM.

L'utente carica un documento, che viene archiviato in Amazon S3. Una funzione AWS Lambda avvia la trasformazione del file, inviandolo a Amazon Textract per l'OCR e l'estrazione del testo. Il documento trasformato viene quindi salvato in un bucket S3 dedicato.

Un'altra funzione Lambda genera embedding vettoriali del testo, utilizzando un modello specifico ospitato su Amazon Bedrock o SageMaker. Gli embedding vengono poi archiviati in Amazon OpenSearch, che funge da database vettoriale per la ricerca semantica.

Le richieste utente vengono gestite tramite Amazon SQS, che accoda le query per una successiva elaborazione asincrona. Una funzione Lambda esegue il retrieval dei documenti rilevanti dal database vettoriale e costruisce un prompt ottimizzato per il modello LLM. Il modello, basato su DeepSeek R1 fine-tuned, esegue l'inferenza e restituisce una risposta contestualizzata all'utente.

I pesi dei modelli personalizzati sono gestiti in Amazon SageMaker Model Registry e archiviati in Amazon S3, garantendo versionamento e tracciabilità delle modifiche. L'uso di Amazon Bedrock per l'inferenza offre un'integrazione scalabile con i modelli fundation, migliorando la flessibilità dell'architettura.

Questa pipeline consente una gestione efficiente dei documenti, ottimizzando il recupero delle informazioni e la generazione delle risposte con LLM, bilanciando scalabilità, costi e performance.

Il trattamento dei documenti avviene in un ambiente controllato, riducendo il rischio di esposizione e garantendo il rispetto dei principi di Privacy by Design.

Tutti i dati vengono crittografati sia a riposo che in transito. I documenti caricati su Amazon S3 sono protetti da crittografia AES-256, mentre le comunicazioni tra i servizi avvengono tramite TLS 1.3, prevenendo intercettazioni e accessi non autorizzati. Anche gli embedding generati e archiviati in Amazon OpenSearch seguono lo stesso principio di protezione, assicurando che le informazioni sensibili non siano esposte.

L'accesso ai dati è strettamente regolato da IAM Policies e autenticazione basata su AWS Cognito, con token JWT che validano l'identità degli utenti prima di consentire operazioni sui documenti. I bucket S3 sono configurati con policy di accesso restrittive, limitando la visibilità dei file solo ai servizi autorizzati. Per garantire la cancellazione dei dati su richiesta, un meccanismo automatico consente agli utenti di eliminare documenti e relativi embedding in conformità con il diritto all'oblio previsto dal GDPR.

Le operazioni su dati e modelli sono monitorate costantemente. AWS CloudTrail registra ogni accesso e modifica ai file, consentendo la tracciabilità e la verifica dell'attività.

Gli aspetti critici relativamente l'uso di queste tecnologie sono:

- Il data leakage, si verifica quando un modello AI restituisce informazioni che non dovrebbero essere accessibili a un determinato utente, spesso a causa di un'errata gestione del contesto o dei dati di addestramento.
- La cross-session contamination, si verifica quando informazioni da una sessione utente vengono involontariamente mantenute e riutilizzate in una successiva, esponendo dettagli di un'interazione precedente a un altro utente.

Queste problematiche sono particolarmente rischiose in scenari multiutente, poiché violano il principio di isolamento dei dati e possono comportare esposizione di dati personali o sensibili, compromettendo la conformità al GDPR.

Soluzioni per mitigare i rischi:

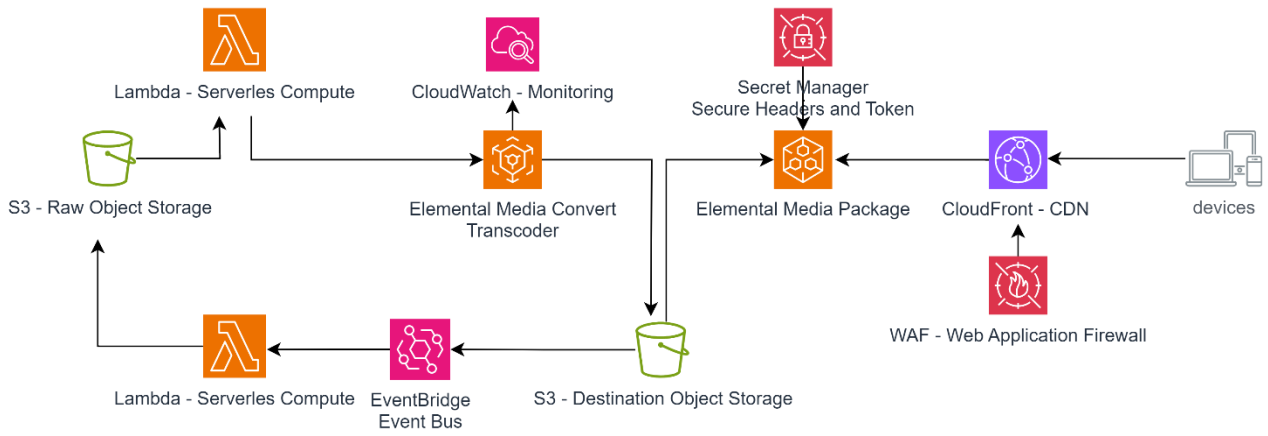
- Isolamento del contesto tra richieste.
    - I modelli AI devono essere stateless, elaborando ogni richiesta indipendentemente da quelle precedenti.
    - Non memorizzare contesti persistenti tra sessioni.
    - Costruire il prompt in modo che utilizzi solo informazioni autorizzate per l'utente corrente.
    - Evitare la cache condivisa per il contesto conversazionale.
- Gestione sicura degli embedding vettoriali.
    - E' essenziale garantire che ogni utente acceda solo ai propri dati.
    - Associare un identificativo univoco (es. User ID) a ogni embedding.
    - Applicare filtri di accesso basati sull'utente prima di eseguire le query sul database vettoriale.

- Controllo degli input e degli output del modello
    - L'input al modello deve essere filtrato per evitare l'inserimento involontario di dati sensibili.

    - Rimuovere riferimenti a sessioni precedenti nel prompt.
    - Implementare validazioni sulle risposte, applicando regex o AI-based filters per intercettare contenuti indesiderati prima della restituzione all'utente.
- Eliminazione forzata dei dati di sessione
    - Dopo ogni richiesta, tutti i dati temporanei devono essere cancellati per prevenire persistenza non intenzionale.
    - Se si utilizza AWS Lambda, evitare di mantenere variabili globali tra esecuzioni.
    - Per sistemi con caching (Amazon ElastiCache/Redis), impostare una TTL (Time-To-Live) breve o cancellare esplicitamente i dati al termine della sessione.
- Logging e auditing delle richieste.
    - Monitorare e analizzare gli input e output dei modelli aiuta a identificare eventuali anomalie.
    - Utilizzare AWS CloudTrail per tracciare le richieste AI.

Implementando un'architettura stateless, con isolamento dei dati, controllo dei prompt e gestione attenta degli embedding, è possibile mitigare il rischio di data leakage e cross-session contamination. L'adozione di controlli di accesso rigorosi e un sistema di monitoring continuo garantisce la protezione delle informazioni sensibili e la conformità al GDPR.

## Livello media




Il processo di ingest e transcodifica dei video segue una pipeline strutturata che garantisce la conversione efficiente dei file sorgenti in formati di streaming adattivo.

- Amazon S3 (Bucket Source). Memorizza i file video caricati dagli utenti. Contiene i file JSON con le specifiche di transcodifica per MediaConvert.
- AWS Lambda (Job Submit). Attivato automaticamente da un evento S3 (upload di un nuovo file). Invia un job ad AWS Elemental MediaConvert con i parametri di transcodifica.
- AWS Elemental MediaConvert. Converte i file sorgenti in HLS, DASH e CMAF Adaptive Bitrate (ABR). Scrive i segmenti e i manifest di output in un bucket Amazon S3 dedicato.
- Amazon CloudWatch + Amazon EventBridge. Monitora i job di transcodifica in MediaConvert. EventBridge rileva il completamento del job e attiva una funzione Lambda.
- AWS Lambda (Job Complete). Gestisce i file generati e aggiorna i metadati. Invia una notifica tramite Amazon SNS (Simple Notification Service) per segnalare la disponibilità del contenuto.

Per proteggere i contenuti e prevenire accessi non autorizzati, l'architettura integra un sistema di gestione degli accessi basato su AWS WAF e AWS Secrets Manager.

- AWS WAF (Web Application Firewall). Configurato su Amazon CloudFront per bloccare traffico malevolo e attacchi DDoS. Implementa policy per limitare l'accesso in base agli IP o alle richieste sospette.
- AWS Secrets Manager. Archivia token di accesso, API keys e header personalizzati. Utilizzato da AWS Lambda per autenticare e autorizzare le richieste di accesso ai contenuti. Impedisce l'accesso diretto ai manifest HLS/DASH senza autenticazione.

L'integrazione tra AWS Elemental MediaPackage e Amazon CloudFront consente una distribuzione globale ottimizzata con caching intelligente e protezione avanzata dei contenuti.

- Amazon S3 (Bucket Destination). Archivia i file transcodificati generati da MediaConvert. Configurato con policy di accesso controllato per MediaPackage.
- AWS Elemental MediaPackage (Opzionale). Converte i segmenti HLS/DASH su richiesta per ridurre lo storage necessario. Applica protezione DRM (Digital Rights Management) per evitare il download non autorizzato. Implementa Just-in-Time Packaging, consentendo una gestione dinamica dei manifest senza doverli pre-generare.
- Amazon CloudFront. Distribuisce i contenuti con caching globale per ridurre la latenza. Serve i contenuti direttamente da MediaPackage o da Amazon S3, a seconda della configurazione scelta. Configurato con Origin Access Identity (OAI) per impedire l'accesso diretto ai file S3. Implementa autenticazione tramite signed URLs o signed cookies per controllare l'accesso.

In alternativa, è possibile evitare l'uso di AWS Elemental MediaPackage e configurare direttamente Amazon S3 come origine per CloudFront.

- Maggiore semplicità: Elimina la necessità di gestire MediaPackage e semplifica la configurazione.
- Riduzione dei costi: Evita il costo aggiuntivo derivante dall'utilizzo di MediaPackage.
- Distribuzione più veloce: CloudFront può servire direttamente i manifest e i segmenti video già presenti in S3 senza ulteriore processing.

Svantaggi di usare solo Amazon S3 come origine:

- Maggior consumo di storage: Poiché i manifest HLS/DASH devono essere generati e archiviati in S3, lo spazio di archiviazione richiesto potrebbe aumentare.
- Assenza di Just-in-Time Packaging: Non è possibile creare manifest HLS/DASH dinamicamente, quindi ogni variante del video deve essere pre-generata.
- Mancanza di DRM nativo: A differenza di MediaPackage, S3 non supporta direttamente la protezione DRM, richiedendo soluzioni di terze parti per la gestione della sicurezza dei contenuti.

Flusso Complessivo dell'Architettura:

- L'utente carica un video su Amazon S3.
- AWS Lambda avvia un job in MediaConvert.
- MediaConvert transcodifica il file e lo salva in S3.
- Opzione 1: MediaPackage preleva i segmenti e genera i manifest HLS/DASH in modo dinamico.
- Opzione 2: CloudFront serve direttamente i file HLS/DASH memorizzati in S3.
- AWS WAF e Secrets Manager proteggono l'accesso ai contenuti.
- CloudFront distribuisce globalmente il video ottimizzando le performance e la sicurezza.
- L'utente finale accede al contenuto con le policy di autenticazione definite.

## Conclusioni

La piattaforma di e-learning ZenClass rappresenta un'infrastruttura avanzata, progettata per garantire un equilibrio tra innovazione, scalabilità e sicurezza. L'integrazione di intelligenza artificiale (NLP, LLM, RAG) con

i servizi cloud di AWS ha permesso la creazione di un ecosistema didattico interattivo, con funzionalità avanzate di gestione dei contenuti, autenticazione sicura e accesso ai dati. Uno degli aspetti fondamentali della piattaforma è l'attenzione alla sicurezza e alla conformità normativa, in particolare rispetto al GDPR e alle best practice ISO 27001. L'adozione di una strategia Zero Trust ha permesso di implementare controlli granulari sugli accessi e sulla gestione dei dati sensibili.

Le principali misure di sicurezza adottate includono:

- Gestione degli accessi e autenticazione:
    - Integrazione con Active Directory per il controllo centralizzato degli utenti.
    - IAM con policy dettagliate per limitare l'accesso ai dati e ai servizi AI.
    - MFA e autenticazione federata con AWS Cognito.
- Protezione della rete e dei servizi cloud:
    - Segmentazione dell'infrastruttura su VPC dedicate per isolare i vari livelli della piattaforma.
    - Protezione da attacchi DDoS e minacce web tramite AWS WAF e AWS Shield.
- Protezione dei dati e conformità GDPR:
    - Crittografia dei dati a riposo e in transito con AWS KMS e TLS 1.3.
    - Monitoraggio e auditing con AWS CloudTrail e AWS Security Hub per garantire tracciabilità degli accessi.
    - Meccanismi di anonimizzazione e pseudonimizzazione per la gestione dei dati sensibili.
- Sicurezza dell'intelligenza artificiale:
    - Prevenzione del data leakage e isolamento del contesto tra sessioni per evitare contaminazioni di dati tra utenti.
    - Controllo degli input/output dei modelli AI per filtrare informazioni sensibili.
    - Logging e auditing delle richieste AI per rilevare anomalie e prevenire usi impropri.

L'infrastruttura ZenClass si distingue per un'implementazione scalabile e resiliente, che sfrutta le potenzialità del cloud per garantire un servizio affidabile e sicuro agli utenti. Le best practice di sicurezza adottate garantiscono la protezione dei contenuti didattici e delle informazioni personali, riducendo al minimo il rischio di violazioni e migliorando l'esperienza complessiva di professori e studenti.

Grazie a una gestione proattiva della sicurezza e un'architettura ottimizzata, ZenClass si posiziona come una piattaforma di e-learning all'avanguardia, capace di rispondere alle esigenze educative moderne senza compromettere la protezione dei dati e la conformità normativa.