Via Nairobi 40
00144, Roma (RM)
info@zendata.it

Sommario

Via Nairobi 40
00144, Roma (RM)
info@zendata.it

Chi siamo

ZenData è una startup dinamica e innovativa, specializzata nello sviluppo di soluzioni basate sull'intelligenza artificiale, per aumentare l'efficienza e stimolare la crescita

Le nostre soluzioni abbracciano una vasta gamma di settori e reparti, spaziando dallo sviluppo di assistenti virtuali per pubbliche amministrazioni e professionisti autonomi, fino all'analisi e all'ottimizzazione dei processi produttivi aziendali.

La nostra strategia è orientata alla progettazione di soluzioni che si integrano in maniera semplice e fluida nei processi operativi dei nostri clienti, garantendo un'adozione senza attriti e un impatto positivo immediato.

Siamo convinti che il nostro approccio all'innovazione, unito alla capacità di personalizzare le soluzioni in base alle specifiche esigenze dei nostri clienti, ci permetta di offrire un valore aggiunto significativo e di contribuire al successo dei nostri partner commerciali.

Via Nairobi 40
00144, Roma (RM)
info@zendata.it

WWW.ZENDATA.IT

1. Introduzione al progetto

1.1 Panoramica

Il presente progetto prevede l'implementazione di **Zenclass** in modalità **SaaS** come iniziativa pilota presso l'**Università Campus Bio-Medico di Roma (UCBM)**.

L'obiettivo è offrire a docenti e studenti una piattaforma evoluta, in grado di analizzare e comprendere contenuti didattici di varia natura (audio, video e documenti), fornendo al contempo un assistente virtuale avanzato per un supporto personalizzato e contestuale.

L'intera soluzione sarà implementata su tecnologie **cloud AWS**, garantendo scalabilità, sicurezza e conformità alle normative vigenti in materia di protezione dei dati.

1.2 Funzionalità

Per i docenti:

- 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 in diverse lingue.
    - Estrazione del contenuto dei documenti dei corsi di laurea.
    - Transcodifica dei video per lo streaming.
    - Archiviazione ottimizzata per la gestione dei contenuti grezzi e processati.

Per gli studenti:

- Accesso autenticato tramite Active Directory, garantendo un controllo centralizzato degli accessi.
- Possibilità di consultare materiali didattici sotto forma di video, documenti.
- Interazione con un piattaforma di Generative AI e con un Chatbot basato su modelli di intelligenza artificiale, in grado di fornire approfondimenti personalizzati sui contenuti dei corsi ed un'esperienza di apprendimento migliorata.

1.3 Perimetro di progetto

L’implementazione del progetto pilota si focalizza su un perimetro limitato, in particolare durante questa fase pilota Zenclass verrà integrato sui seguenti corsi:

- **Elaborazione segnali**, previsto nel terzo anno del corso di Laurea Triennale di Ingegneria Industriale. Il materiale previsto da integrare in Zenclass include:
    - dispense in formato Markdown;
    - video per la durata complessiva di 60 ore.
- **Architettura dei sistemi distribuiti**, previsto nel primo anno del corso di Laurea Magistrale di Ingegneria dei Sistemi Intelligenti. Il materiale previsto da integrare in Zenclass include:
    - lucidi in formato pptx;
    - dispense in formato MarkDown.

1.4 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

- 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.

2. Architettura

La seguente sezione descrive l'architettura della piattaforma Zenclass, suddivisa in più livelli funzionali per garantire modularità e sicurezza. In particolare, sono previsti i seguenti livelli:

- **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.

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.

2.1 Livello di autenticazione

Il livello di autenticazione prevede 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:

1. L'utente accede alla piattaforma ZenClass tramite l'interfaccia web o mobile.

2. L'applicazione frontend reindirizza l’utente verso AWS Cognito, che funge da Identity Broker.
3. 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:

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

Autenticazione tramite ADFS:

6. L’utente inserisce le credenziali aziendali (username e password di Active Directory).
7. ADFS interroga Active Directory (AD) per verificare l’identità dell’utente.
8. 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.

2.2 Livello applicativo

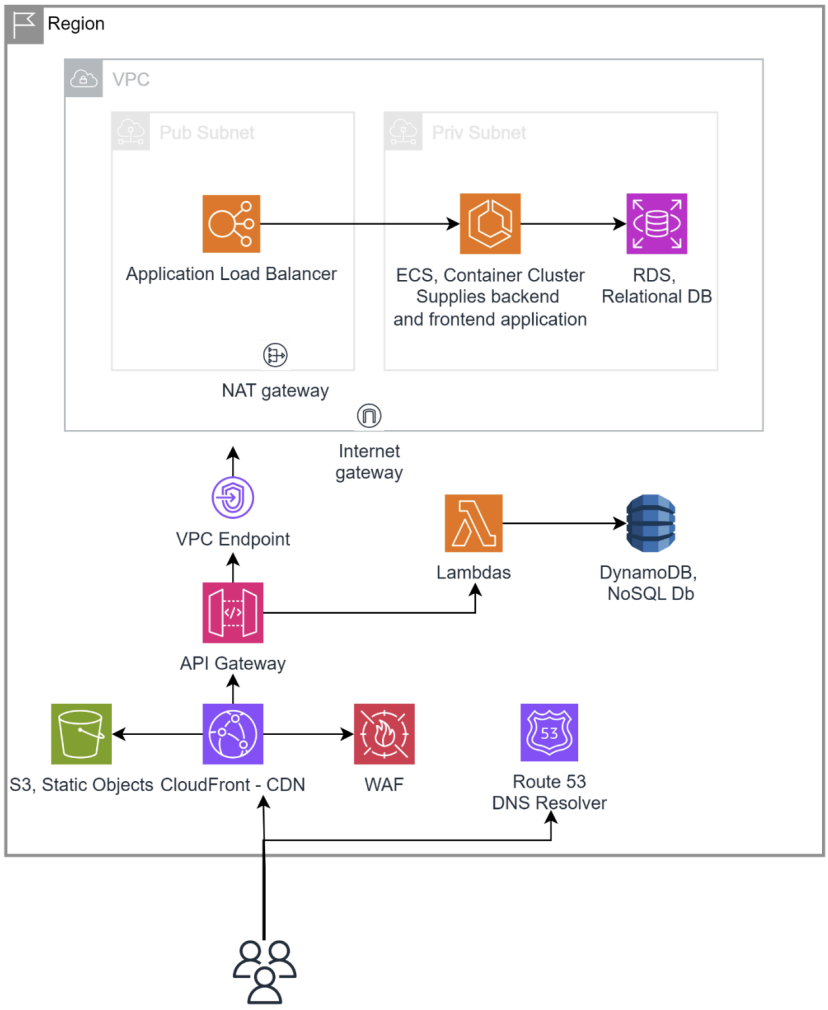
L'architettura applicativa della piattaforma ZenClass è progettata per gestire richieste utente in modo efficiente, distribuire il carico di lavoro e garantire sicurezza e scalabilità. Il flusso delle richieste si articola attraverso diversi servizi AWS, che interagiscono tra loro per offrire una risposta 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 ad esempio 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**, **CloudTrai**l assicurando il rilevamento di anomalie e la tracciabilità degli accessi.



2.3 Livello AI

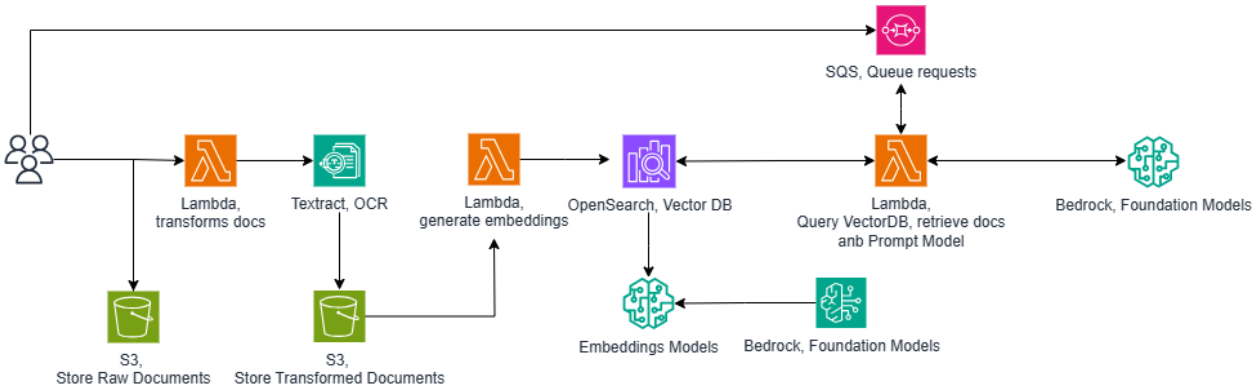
L'architettura AI della piattaforma ZenClass è progettata per gestire l'elaborazione documentale, l'indicizzazione vettoriale e l'inferenza tramite modelli di linguaggio. Il flusso dei dati segue un processo strutturato che trasforma documenti grezzi in rappresentazioni ottimizzate per il retrieval e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



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.

2.4 Livello Media

Il processo di ingestion e transcodifica dei video viene semplificato grazie a Cloudflare Stream, che integra in un’unica piattaforma le funzioni di upload, transcodifica e distribuzione dei contenuti video in modalità streaming adattivo. Di seguito, il flusso operativo aggiornato:

- **Upload e Ingest:**
  Gli utenti caricano i video direttamente su Cloudflare Stream tramite le API o l’interfaccia web.
- **Transcodifica Automatica:**
  Una volta caricato, Cloudflare Stream gestisce automaticamente la transcodifica del video in più risoluzioni e bitrates. Il servizio genera in automatico i manifest e i segmenti necessari per lo streaming adattivo.
- **Distribuzione Globale:**
  I video transcodificati vengono distribuiti attraverso la rete CDN globale di Cloudflare, garantendo bassa latenza e alta disponibilità senza configurazioni aggiuntive. Il sistema supporta meccanismi di caching intelligente integrati per ottimizzare le performance.

Flusso Complessivo dell’Architettura con Cloudflare Stream:

- L’utente carica un video direttamente su Cloudflare Stream.
- Il servizio esegue la transcodifica automatica, creando manifest e segmenti per lo streaming adattivo.

- I contenuti vengono distribuiti globalmente tramite la rete CDN di Cloudflare, con performance ottimizzate e sicurezza integrata.

2.5 Security

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.

3. Struttura dell’offerta

Di seguito viene presentata la stima dei costi di progetto:

Costi di Progetto	
Servizi professionali Zendata	
implementazione (CapEx)	10.000 €
manutenzione (OpEx) (esclusa infrastuttura)	1.000 €/mese

Tutti i prezzi sono da considerarsi IVA esclusa.

Di seguito si riportano i costi in dettaglio dell’infrastruttura:

Nome Servizio	Descrizione	€/mese
Cloudflare Stream	Piattaforma di streaming video on demand	160
S3	Servizio di archiviazione oggetti che offre scalabilità, disponibilità dei dati e sicurezza	10
WAF	Servizio che protegge applicazioni web da attacchi comuni come SQL injection, cross-site scripting (XSS) e attacchi DDoS	8
DynamoDB	Db NoSql per la gestione delle conversazioni	20
RDS	Db Sql per la gestione dei dati strutturati	20

ECS	Applicativo Backend e Server Side front end	100
Load Balancer	Servizio per la equidistribuzione delle richieste al layer applicativo	22
Opensearch	Database Vettoriale per Information retrieval del RAG System	75
Extra (KMS, Secret Manager, Cloudwatch, Lambda)	**KMS**: Servizio per la gestione e crittografia delle chiavi **Secrets Manager**: Archivia e gestisce in modo sicuro credenziali, API keys **CloudWatch**: Monitoraggio e log delle applicazioni **Lambda**: Serverless function per gestione delle richieste applicative	100
Totale		515

Il dimensionamento delle macchine è stato volutamente sovrastimato in quanto il reale consumo è legato al numero di richieste / traffico sul portale e per tale motivo può fare uso di meccanismi di upscale e downscale che permettono dunque di ridurre l'over allocazione delle risorse cloud e consumi contenuti.

Di seguito è riportata anche una stima dei costi per i servizi del livello di AI. Questa stima considera un volume di 23.000 messaggi/mese, basata sull'ipotesi che, in un arco di 6 mesi, tutti gli studenti completino l'intero corso di loro competenza e interagiscano con il sistema con una frequenza media di 15 domande all'ora. I dettagli sono i seguenti:

Nome Servizio	Descrizione	€/mese
AI Services	Large Language Models	230

AI Services	Embedding Models	2
Speech-to-Text	Speech-to-Text per 23k messaggi/mese (tutti quelli inviati)	23
Text-to-Speech	Text-to-Speech per 23k messaggi/mese (tutti quelli ricevuti)	276
Totale		531

Per i servizi di Speech-to-Text e Text-to-Speech, si è volutamente assunto che vengano utilizzati su tutti i messaggi inviati e ricevuti, al fine di fornire una valutazione completa dei costi potenziali.

Tutte le stime di costo sono basate su ipotesi iniziali, il costo scalerà in proporzione al numero di messaggi effettivi in produzione.

La licenza è considerata valida fino al 30 settembre 2025.

zendata

info@zendata.it

www.zendata.it

Via Nairobi 40
00144, Roma (RM)
info@zendata.it

zendata

WWW.ZENDATA.IT