# OrientAi – Asistente de Orientación Vocacional con Inteligencia Artificial

**Integrantes: Jaime David Mejia Quintero**

**Juan Zuluaga**

**Julian Seohanes**

**Curso: Arquitectura en la Nube – Nivel Innovador**

**Fecha: 06/08/2025**

# Introducción

En el marco del curso de Arquitectura en la Nube, se desarrolló una aplicación web full stack que utiliza inteligencia artificial para asistir a estudiantes en su proceso de orientación vocacional. Esta aplicación permite a los usuarios responder una serie de preguntas personalizadas (en un principio generadas por nosotros) pero en un futuro sería generadas por una IA, que analiza las respuestas y sugiere posibles carreras profesionales según los intereses, habilidades y preferencias del estudiante. Además de esto provee otras funcionalidades (Guardado de la información del estudiante, propuestas en cuanto a universidades y becas que el estudiante puede adquirir según la carrera propuesta, el colegio también puede recopilar información de los estudiantes que accedieron a la plataforma, entre otros). El objetivo principal es brindar una herramienta tecnológica innovadora y accesible que ayude a tomar decisiones más informadas sobre el futuro académico y profesional.

OrientAi no solo ayuda a los estudiantes a tomar decisiones académicas informadas, sino que proporciona a instituciones educativas herramientas valiosas para mejorar sus procesos internos, análisis de datos y seguimiento personalizado del estudiantado. Este documento describe claramente tanto el estado actual como la visión futura del proyecto, abordando sus aspectos técnicos y estratégicos.

# Antecedentes

Actualmente en el mercado existen ciertas soluciones orientadas a la propuesta de vocación de carrera, entre ellas se encuentran las siguientes :

MyNextMove: Plataforma gubernamental de EE.UU. con foco en datos del mercado laboral, sin IA generativa.-

CareerExplorer: Test profundo con más de 800 variables, matching profesional detallado, sin integración institucional.-

Chatbot IA: Soluciones flexibles con GPT/Claude, sin estructura persistente de datos o exportación.

# Cuadro comparativo

| Nombre de la solución | Características principales | Uso de IA generativa | Integración institucional | Persistencia de datos / exportación |
|---|---|---|---|---|
| MyNextMove | Plataforma gubernamental de EE.UU. enfocada en datos del mercado laboral | No | No | Limitada |
| CareerExplorer | Test vocacional profundo con más de 800 variables y matching profesional detallado | No | No | Sí (limitada al usuario) |
| Chatbot IA (GPT/Claude) | Soluciones flexibles basadas en IA generativa como GPT o Claude | Sí | No | No |

¿Entonces que propone orientAi que sea diferente a estas propuestas ?

- Exportación de resultados a archivos PDF que estarán disponibles para descargar posteriores
- Dashboards para uso institucional
- Apoyo de entrenamiento de IA con sagemaker o in-house para dar mejores resultados

# Desarrollo del Proyecto

## 1. Arquitectura de la Aplicación:

- Frontend: Desarrollado en React.js

- Backend: Node.js con Express

- Base de Datos: Mysql (Pensando en migrar a la nube hacia una dynamoDB o Aurora)

- IA: Ollama integrado mediante API para generar preguntas y sugerencias de carrera

- Despliegue: Se basó localmente en kubernetes, se quiere utilizar nube contando con servicios de AWS (EC2, S3, CloudWatch, etc.)

- CI/CD: GitHub (Actions)

- Infraestructura como Código: Terraform para la creación y gestión de recursos en la nube

## 1. Justificación del stack tecnológico de orient-AI:

A continuación, se presenta brevemente la justificación de las tecnologías seleccionadas para la arquitectura actual:

- **React.js:**
  Elegido debido a su alto rendimiento, flexibilidad y gran comunidad que permite desarrollos rápidos y mantenibles.

- **Node.js con Express:**
  Seleccionado por su modelo asincrónico que maneja eficazmente operaciones simultáneas, facilitando APIs rápidas y escalables, ideal para entornos ágiles.

- **MySQL (migración futura a DynamoDB/Aurora):**
  Actualmente es usado por su estabilidad y familiaridad. La futura migración a DynamoDB o Aurora proporcionará mejor escalabilidad, rendimiento, y flexibilidad en la nube.

- **IA (Ollama vía API):**
  Elegido debido a su privacidad, control de datos locales, reducción de costos operativos, y flexibilidad frente a alternativas externas como OpenAI o Vertex AI.

- **Kubernetes local (AWS futuro):**
  Facilita la portabilidad y escalabilidad inicial, que permitirá una transición fluida hacia AWS ECS/Fargate.

- **GitHub Actions:**
  Integrado por simplicidad en el manejo de pipelines de CI/CD y la integración natural con GitHub.

- **Terraform:**
  Seleccionado por su capacidad multicloud y uso amplio en la industria para automatizar infraestructura, facilitando el mantenimiento a largo plazo.

## Cuadro comparativo Desarrollo de solución

| **Componente** | **Tecnología Actual (Propuesta)** | **Alternativas comunes** | **Ventajas de la elección actual** | **Consideraciones de las alternativas** |
|---|---|---|---|---|
| Frontend | React.js | Vue.js / Angular / Python con Flask front | Amplia comunidad, rápida, altamente integrable con componentes modernos | Vue es más simple pero menos robusto para apps complejas, Angular es más pesado |
| Backend | Node.js con Express | Java con Spring Boot / Python con Django | Asincronía nativa, desarrollo ágil, buen rendimiento para APIs | Java más robusto pero verboso; Python más fácil pero menos eficiente en concurrencia |
| Base de Datos | MySQL (migrando a DynamoDB o Aurora) | PostgreSQL / MongoDB | Estructurada, madura; DynamoDB/Aurora son altamente escalables en AWS | PostgreSQL tiene mejor manejo de relaciones complejas; MongoDB es más flexible pero menos estructurado |
| IA (Integración) | Ollama vía API | OpenAI / Claude / Vertex AI | IA localizable, control de costos y privacidad de datos | OpenAI y similares ofrecen más capacidades pero a mayor costo o menos control |
| Despliegue | Kubernetes local (planeado en AWS) | Docker Swarm / Serverless (Lambda) | Escalable y portátil, permite mover cargas a la | Swarm más simple pero menos soportado; |

| | | | | |
|---|---|---|---|---|
| | | | nube fácilmente | Lambda reduce costos pero limita control |
| CI/CD | GitHub Actions | Jenkins / GitLab CI | Integración nativa con GitHub, simple y sin servidores adicionales | Jenkins potente pero complejo de mantener; GitLab CI más cerrado |
| Infraestructura como Código | Terraform | Pulumi / AWS CloudFormation | Multicloud, declarativo, ampliamente adoptado por DevOps | Pulumi permite usar lenguajes conocidos; CloudFormation es más rígido y solo AWS |

## 2. Funcionalidad Principal de la aplicación Orient-AI:

- Registro y autenticación de usuarios(tanto estudiantes, profesores y administradores de las plataformas)

- Cuestionario personalizado de orientación vocacional

- Procesamiento de respuestas mediante IA

- Generación de recomendaciones de carrera en tiempo real

- Historial de resultados y sugerencias

# Resultados Esperados

Se hace la implementación inicial de una aplicación web en la cual un estudiante por medio de una interfaz logra registrarse, se le hace una serie de preguntas y al final la IA determina una sugerencia de una posible carrera a estudiar según sus preferencias, adicionalmente puede descargar el PDF para guardar su resultado, esto es hecho “localmente” para eso se utilizaron herramientas como docker, kubernetes , manejadores de paquetes como NPM y herramientas de monitoreo y seguridad aprendidas en el curso, **¿ Que se espera para futuro?**. Mudar todo a la nube, para esto se tiene la siguiente arquitectura :

# Proximos pasos

Se espera a futuro mudar todo a la nube, para esto ya se tiene una arquitectura planeada que consta de los siguientes elementos :

**Infraestructura de Red**

- **VPC (Virtual Private Cloud)**: Red privada virtual que proporciona aislamiento de red
- **Subredes Públicas**: Para componentes que requieren acceso directo desde Internet
- **Subredes Privadas**: Para componentes internos que no necesitan exposición directa
- **NAT Gateway**: Permite conectividad saliente para recursos en subredes privadas
- **VPC Endpoints**: Conexión privada a servicios de AWS sin tráfico de Internet

**Capa de Aplicación**

- **Application Load Balancer (ALB)**: Distribuye el tráfico entrante entre múltiples instancias
- **Amazon ECS (Elastic Container Service)**: Orquestación de contenedores
- **AWS Fargate**: Plataforma serverless para ejecutar contenedores sin gestionar servidores
- **Auto Scaling Group**: Escalado automático basado en demanda
- **Docker**: Containerización de la aplicación api-orientAi

**Base de Datos y Almacenamiento**

- **Amazon DynamoDB**: Base de datos NoSQL para almacenar:
    - Información de estudiantes
    - Respuestas del cuestionario

    - Banco de preguntas
- **Amazon S3**: Almacenamiento de objetos para archivos estáticos y backups

**Inteligencia Artificial Amazon SageMaker**:

    - Entrenamiento de modelos de IA
    - Inferencia para recomendaciones vocacionales
    - Procesamiento de respuestas del cuestionario

**Seguridad y Gestión de Configuración**

- **AWS IAM (Identity and Access Management)**: Gestión de identidades y permisos
- **AWS Identity Center**: Centralización de acceso para usuarios
- **AWS Secrets Manager**: Gestión segura de credenciales y secretos
- **Parameter Store**: Almacenamiento de parámetros de configuración

**Monitoreo y Observabilidad**

- **Amazon CloudWatch**: Monitoreo de métricas, logs y alertas
- **Grafana**: Dashboard personalizado para visualización de métricas
- **Alertas automatizadas**: Notificaciones en caso de fallos o anomalías

**DevOps y Automatización**

- **AWS CloudFormation**: Infraestructura como código (IaC)
- **GitHub Actions**: Pipeline de CI/CD para despliegue automatizado

**Flujo de Datos**

1. **Acceso del Usuario**: Los estudiantes acceden a la aplicación a través del Application Load Balancer
2. **Procesamiento**: Las solicitudes son dirigidas a los contenedores en ECS Fargate
3. **Almacenamiento**: Los datos se almacenan de forma segura en DynamoDB
4. **IA Processing**: Las respuestas son procesadas por SageMaker para generar recomendaciones
5. **Respuesta**: Los resultados son devueltos al usuario en tiempo real

# ESTRATEGIAS

## Estrategia Integral de Escalabilidad y Rendimiento

La aplicación OrientAi será capaz de manejar incrementos significativos en tráfico mediante la implementación de:

- **AWS ECS con Auto-Scaling y Fargate:**
  Permite escalar automáticamente según la demanda, optimizando costos y manteniendo un rendimiento consistente.

- **Application Load Balancer:**
  Distribuirá eficientemente el tráfico para garantizar tiempos rápidos de respuesta y alta disponibilidad.

- **Bases de datos NoSQL (DynamoDB):**
  Garantizan una alta escalabilidad horizontal con baja latencia.

## Estrategia Integral de Seguridad

La seguridad se implementará integralmente en todas las capas de la arquitectura:

- **Infraestructura:**
  Uso de VPC privadas, IAM roles, políticas estrictas, y monitoreo continuo mediante CloudWatch.

- **Aplicación:**
  Implementación robusta de autenticación, autorización, control de sesiones seguras y encriptación TLS.

- **Datos:**
  Uso de cifrado en reposo y tránsito (S3, DynamoDB), cumplimiento de regulaciones internacionales y nacionales (GDPR, protección de datos personales).

## Implementación efectiva de DevOps y IaC

- **Pipeline CI/CD con GitHub Actions:**
  Integrará pruebas automatizadas, análisis de seguridad estático, creación de imágenes Docker, y despliegues automáticos a AWS.

- **Gestión de versiones y rollback:**
  Uso de versionamiento semántico y configuración clara en ECS para facilitar

rollback inmediato en caso de problemas críticos.

- **Terraform:**
  Uso consistente para manejar toda la infraestructura como código, proporcionando transparencia, reproducibilidad y facilidad en el mantenimiento.

## Inteligencia Artificial (SageMaker)

La estrategia para la implementación de la IA será:

- **Entrenamiento de modelos (SageMaker):**
  Se recolectarán inicialmente datos internos anonimizados para entrenar modelos específicos orientados a la generación dinámica de cuestionarios vocacionales.

- **Comparativa de soluciones:**
  Ollama ofrece ventajas significativas en privacidad y control local sobre alternativas como OpenAI y Vertex AI, justificando claramente su uso desde una perspectiva estratégica y técnica.

## Análisis Preliminar de Costos en AWS

Se realizará un análisis inicial sobre costos esperados considerando:

- Uso optimizado de AWS Fargate y autoescalado para reducir costos operativos.
- Uso estratégico de instancias reservadas para ahorro económico significativo.
- Optimización del almacenamiento en DynamoDB y S3 con políticas de ciclo de vida.

| Servicio AWS | Descripción/uso | Unidad/Cantidad | Costo mensual estimado (USD) |
|---|---|---|---|
| **Amazon ECS (AWS Fargate)** | 2-4 tareas simultáneas, uso moderado (~vCPU 0.5-1GB RAM c/u) | ~720 horas al mes (24/7) | ~$35 – $50 |

| | | | |
|---|---|---|---|
| **Application Load Balancer (ALB)** | Balanceo de carga para alta disponibilidad | 1 ALB, tráfico moderado | ~$20 |
| **Amazon DynamoDB** | Almacenamiento NoSQL, hasta 10GB | Hasta 10 GB, lecturas/escrituras moderadas | ~$15 – $25 |
| **Amazon S3** | Almacenamiento de resultados PDF y archivos estáticos | 50 GB almacenamiento, tráfico bajo | ~$5 – $10 |
| **Amazon SageMaker** | Entrenamiento e inferencia moderada | 20 horas entrenamiento mensual, inferencias bajo demanda | ~$50 – $70 |
| **Amazon CloudWatch** | Monitoreo continuo, logs, métricas | Uso moderado | ~$5 – $15 |
| **AWS Secrets Manager** | Gestión segura de credenciales | Hasta 10 secretos administrados | ~$5 – $10 |
| **NAT Gateway** | Conectividad saliente desde subredes privadas | 1 Gateway, tráfico moderado | ~$35 – $40 |
| **Costos de red (Transferencia)** | Datos transferidos, ~50 GB | 50 GB | ~$5 – $8 |

| | | | |
|---|---|---|---|
| **AWS IAM y AWS Identity Center** | Gestión de usuarios y accesos | Sin costo adicional | $0 |
| **AWS CloudFormation/Terraform** | Infraestructura como Código | Sin costo adicional | $0 |
| **GitHub Actions (CI/CD)** | Despliegues automatizados | Uso moderado (CI/CD básico) | $0 (gratis hasta uso moderado) |
| **Amazon VPC, Subredes, Endpoints** | Infraestructura básica de red | Sin costo adicional (solo endpoints) | ~$5 – $10 |
| **Contingencia (10%)** | Posibles variaciones no previstas | - | ~$20 – $25 |
| **Total mensual estimado** | | | **~$195 – $293 USD** |

## Gestión y Mitigación de Riesgos Técnicos

- **Disponibilidad y recuperación rápida:**
  Uso de arquitectura de alta disponibilidad, balanceo de cargas, redundancia regional (AWS).

- **Seguridad:**
  Auditorías periódicas, monitoreo continuo de eventos de seguridad y políticas estrictas de gestión de vulnerabilidades.

- **Integridad y disponibilidad de datos:**
  Política estricta de backup automatizado (AWS Backup), recuperación rápida ante incidentes, y gestión estricta del acceso y permisos.

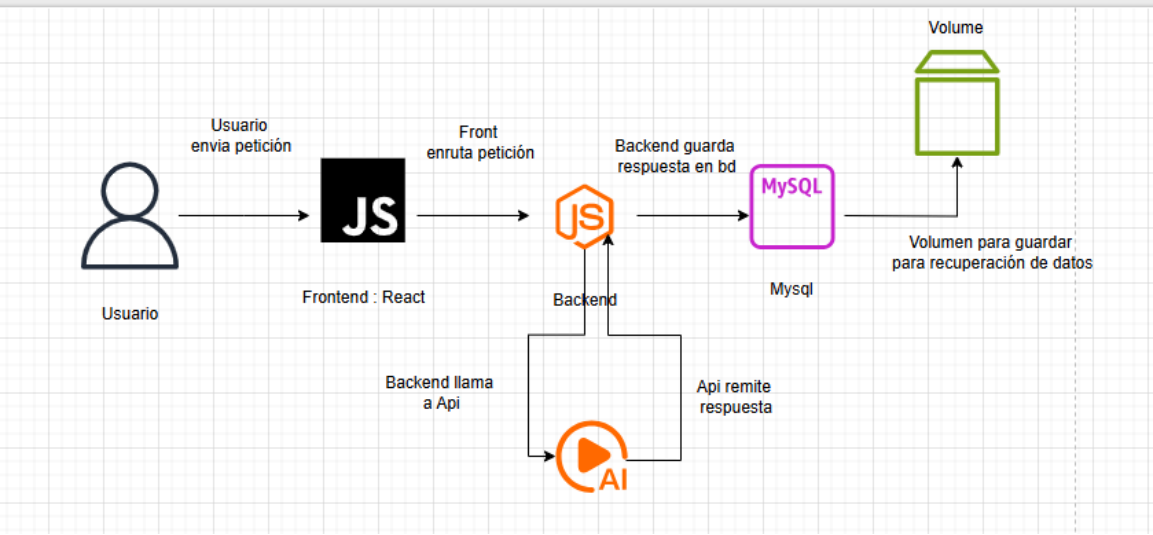
## Representación gráfica de la arquitectura de orient-AI

Para facilitar el entendimiento de la arquitectura actual y propuesta, se incluirán diagramas UML que muestran claramente componentes tecnológicos específicos (React.js, Node.js, MySQL), servicios AWS, y sus interacciones internas. Cada diagrama estará acompañado de una descripción breve que explica los componentes claves, flujos de información, y medidas de seguridad y escalabilidad integradas.

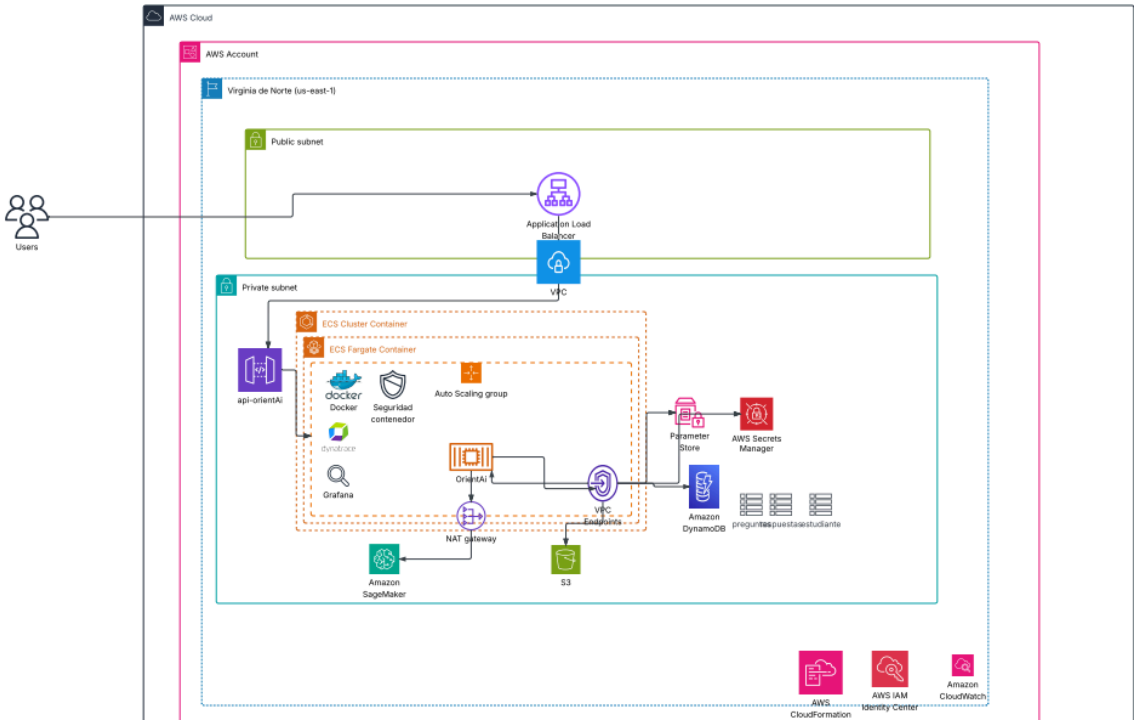
- **Arquitectura Local (actual):** Representación gráfica de los componentes actuales, sus interacciones y limitaciones actuales de escalabilidad y rendimiento.

- **Arquitectura Cloud (futura):** Esquema detallado con los servicios en AWS (ECS, Fargate, DynamoDB, SageMaker, CloudWatch, etc.), claramente diferenciados por capa: red, aplicación, datos, seguridad, monitoreo, y automatización (IaC/DevOps).

## Arquitectura actual (Local)




## Arquitectura próximos pasos


AWS Cloud
AWS Account
Virginia de Norte (us-east-1)
Public subnet
Users
Application Load Balancer
VPC
Private subnet
ECS Cluster Container
ECS Fargate Container
api-orientAi
Docker
Seguridad contenedor
Auto Scaling group
dynatrace
Grafana
OrientAi
Parameter Store
AWS Secrets Manager
VPC Endpoints
Amazon DynamoDB
preguntas
respuestasestudiante
NAT gateway
Amazon SageMaker
S3
AWS CloudFormation
AWS IAM Identity Center
Amazon CloudWatch