**PROYECTO GOOGLE YELP.**

**(HOTELES)**

**Tabla de contenido**

[KPIs 1](#_heading=h.gjdgxs)

[Alcance del proyecto 2](#_heading=h.30j0zll)

[EDA 3](#_heading=h.1fob9te)

[Repositorio en Github 3](#_heading=h.3znysh7)

[Implementación Stack Tecnológico 3](#_heading=h.2et92p0)

[Metodología de trabajo 4](#_heading=h.tyjcwt)

[Diseño detallado 4](#_heading=h.3dy6vkm)

[Fuentes de información 4](#_heading=h.17dp8vu)

[Equipo de trabajo – roles y responsabilidades 4](#_heading=h.1t3h5sf)

[Cronograma general – GANT 4](#_heading=h.4d34og8)

[Análisis preliminar de calidad de datos 4](#_heading=h.2s8eyo1)

# KPIs

| **KPI** | **Descripción** | **Objetivo** | **Fórmula** | **Periodicidad** | **Comentarios** |
| --- | --- | --- | --- | --- | --- |
| **Puntuación Promedio de Satisfacción en Reseñas** | Medir la satisfacción del cliente a través de las reseñas. | Mantener una puntuación de satisfacción de 4.5 estrellas o superior | Promedio de las calificaciones | Anual | Este KPI ayuda a evaluar la salud financiera del hotel. |
| **ROI (Retorno de la inversión)** | Medir la rentabilidad de la inversión en el hotel. | Lograr un ROI del 10% anual | (Ganancia neta / Inversión) \* 100 | Anual | Este KPI ayuda a obtener información sobre la experiencia del cliente y mejorar la satisfacción. |
| **Densidad de Hoteles por Zona** | Número de hoteles por milla cuadrada en áreas clave, lo que indica saturación de mercado. | Mantener la densidad de hoteles por debajo de 5 hoteles por milla cuadrada en áreas clave | Hoteles / Millas cuadradas | Anual | Este KPI ayuda a tomar decisiones estratégicas sobre la expansión del hotel. |

| **Índice de Competitividad** | Comparación de la calificación promedio de un hotel con la calificación promedio de otros hoteles en la misma área. | Superar la calificación promedio de la competencia 0.5 puntos | (Calificación promedio del hotel - Calificación promedio de la competencia) | Trimestral | Este KPI ayuda a medir el posicionamiento del hotel en comparación con la competencia. |
| --- | --- | --- | --- | --- | --- |

| **Tasa de Crecimiento de Reseñas Anuales** | Medir el aumento porcentual año tras año en el número de reseñas para hoteles cercanos a los estadios. | 15% de crecimiento anual | ((#Reseñas este año - #Reseñas año anterior) / #Reseñas año anterior) \* 100 | Anual | Este KPI ayuda a medir la visibilidad del hotel y la satisfacción del cliente. |
| --- | --- | --- | --- | --- | --- |

El objetivo final del proyecto contratado es recomendar los cinco (5) estadios donde pudiera existir la mayor probabilidad de éxito para la inversión respectiva en el negocio de hoteleria.

# 

# Alcance del proyecto

Hoteles y negocios que queden alrededor de los estadios donde se jugarán los partidos del mundial del año 2026 en Estados Unidos. En la Figura 1 se detallan las características de cada estadio.

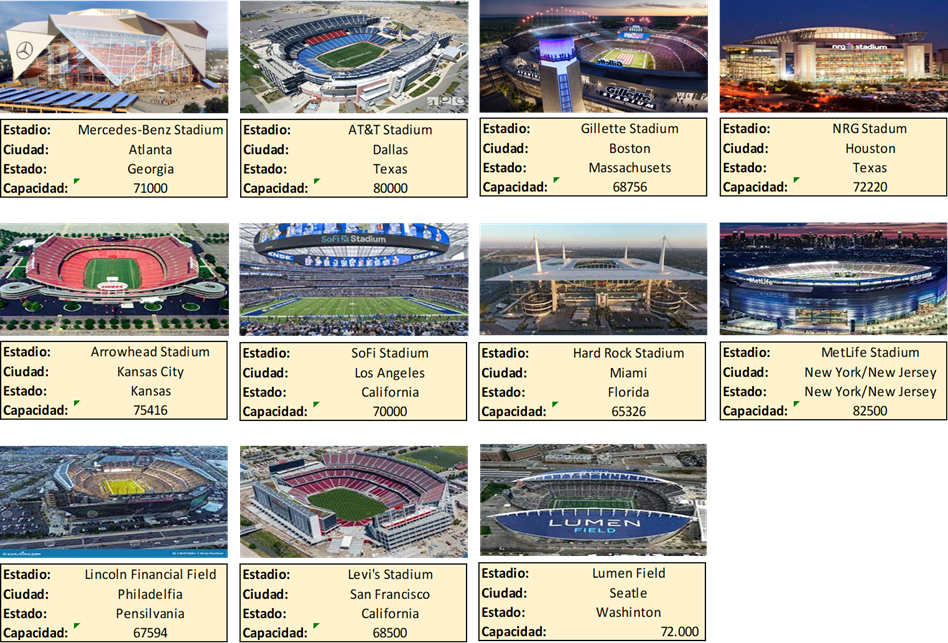


Figura 1. Estadios donde se jugarán los partidos del mundial

Por otro lado, y en vista de la necesidad de los inversionistas en los temas de hospedaje, el proyecto se enfocará en los hoteles y sitios dormitorio en general que se encuentren ubicados en las bases de datos; además porque es necesario también recomendar sitios de estadía a los viajeros.

La empresa de inversiones Adubai Invest, grupo empresarial árabe con sede principal en New York e inversiones en varias ciudades de Estados Unidos y Europa está auscultando la posibilidad de realizar inversiones en hotelería en las ciudades de Estados Unidos sede del mundial de fútbol 2026. Para este propósito contrató la empresa Opportunity Hunters, especialista en buscar oportunidades de negocio mediante el análisis de datos tomados de redes sociales y páginas donde los diferentes usuarios recomiendan sitios de compra.

En este sentido el alcance del proyecto está dado en los siguientes parámetros:

* **Lugares objetivo**: 10 estadios donde se jugarán los partidos del mundial
* **Ciudades objetivo**: 10 ciudades donde están ubicados los estadios respectivos
* **Sitios de interés para el análisis**: todos los puntos referenciados en la metadata de Google y Yelp con ubicación geográfica ‘latitude’ y ‘longitude’ con distancia menor a 30 kilómetros alrededor del estadio respectivo, y se agregarán algunas zonas de interés turístico de acuerdo a la ciudad.

Esta distancia objetivo la determinamos teniendo en cuenta que en Estados Unidos los sitios en general quedan muy distantes y seguramente los visitantes buscarán aquellos lugares de estadía más cercanos al estadio.

Por otro lado, y en vista de la necesidad de los inversionistas en los temas de hospedaje, el proyecto se enfocará en los hoteles y sitios dormitorio en general que se encuentren ubicados en las bases de datos; además porque es necesario también recomendar sitios de estadía a los viajeros.

En la Tabla 1, se detallan los estadios, la ciudad donde está ubicada, la capacidad en personas y el estado.



Tabla 1. Estadios donde se jugarán los partidos del mundial 2026 en Estados Unidos

# 

# EDA

# Repositorio en Github

# Implementación Stack Tecnológico

Para el desarrollo del proyecto, se han elegido cuidadosamente una serie de herramientas tecnológicas que permitirán un procesamiento eficiente y escalable de los datos. A continuación, se presentan las herramientas seleccionadas del Stack Tecnológico:

Python: Utilizado para el procesamiento y análisis de datos

Pandas y Numpy: Librerías para el manejo tabular de datos.

Matplotlib y Seaborn: Herramientas para visualización y análisis exploratorio de datos.

NLTK (Natural Language Toolkit): Librería para procesamiento de lenguaje natural, incluyendo análisis de sentimiento a las reseñas hechas por los usuarios de ambas plataformas.

boto3: Librería para interactuar con servicios de AWS como Amazon S3.

Docker: Sistema para gestionar contenedores, simplificando la manipulación de datos.

Amazon S3 AWS: Almacenamiento escalable y duradero para datos a través de internet.

GitHub: Plataforma de control de versiones para colaboración en equipo.

Google Cloud Storage: Ideal para datasets escalables y optimizados en la nube.

Google Drive: Utilizado para sincronización y acceso a archivos desde cualquier lugar.

Looker Studio: Es una herramienta utilizada para la visualización de datos a través de paneles o informes.

# Metodología de trabajo

Enfoque del Proyecto:

El equipo del proyecto final se ha embarcado en una simulación de un ambiente laboral para fortalecer competencias técnicas y suaves, enfatizando la sinergia y la respuesta a los requerimientos de un Product Owner.

Estructura y Roles:

Se establecieron roles dentro del equipo, asignando responsabilidades principales que abarcan desde la ingeniería de datos hasta la analítica, asegurando que el proyecto reciba una atención detallada en cada fase crítica.

Ciclo de Revisión y Mejora:

Un Head Mentor supervisa diariamente, ofreciendo apoyo general y asesoramiento técnico, aunque la ejecución de tareas recae en la autonomía del equipo.

Metodología Ágil y Herramientas de Planificación:

Adoptando Scrum, el equipo trabaja en sprints, utilizando Jira para gestionar tareas y un diagrama de Gantt para la planificación. Esto permite un seguimiento detallado y la adaptación a cambios durante el proyecto.

Hitos y Entregables:

Los hitos se establecen como indicadores clave de progreso, mientras que los entregables son productos específicos resultantes de las tareas realizadas.

Documentación Continua:

La documentación meticulosa es una práctica central, manteniendo un registro detallado de cada acción y decisión.

Planificación y Herramientas:

El equipo desarrolla un cronograma detallado y selecciona un stack tecnológico que permita abordar eficientemente la limpieza y transformación de datos.

Adelanto y Retroalimentación:

Se anticipan tareas de sprints futuros para obtener retroalimentación temprana del PO y mantener el alineamiento con los objetivos.

Sprints del Proyecto:

Sprint 1: Se enfoca en el análisis inicial y la propuesta de cómo abordar el proyecto, estableciendo un entendimiento claro de la situación y objetivos.

Sprint 2: Se trabaja en la infraestructura del proyecto con un enfoque en ETL, preparando el terreno para análisis y visualizaciones de datos.

Sprint 3: Culmina con el desarrollo de dashboards interactivos y modelos de ML, integrando análisis y datos procesados para proporcionar insights valiosos.

En cada sprint, la iteración, la colaboración y la flexibilidad son fundamentales, guiando al equipo hacia la entrega de soluciones innovadoras de data.

# Diseño detallado

**OPCION HIBRIDA**

Para crear un diseño detallado del proyecto, que integra datos en Parquet y JSON y maximiza los créditos gratuitos de plataformas en la nube, se propone una estructura que abarca desde la ingesta hasta la visualización y monitoreo, con el objetivo de desarrollar un sistema eficiente y costo-efectivo que respete el ciclo de vida del dato:

Ingesta de Datos:

Almacenamiento: Utilizar AWS S3 para guardar los datasets en Parquet y JSON, aprovechando su capa gratuita.

Automatización: Implementar scripts o funciones AWS Lambda que automaticen la ingesta de datos.

Procesamiento y Análisis:

Computación y Desarrollo: Emplear Google Colab para el procesamiento de datos y entrenamiento de modelos de ML, usando sus GPUs gratuitas.

Análisis a Gran Escala: Adoptar Google BigQuery para el análisis de datos, aprovechando su cuota gratuita.

Despliegue y Orquestación:

Automatización de Workflow: AWS Step Functions para gestionar flujos de trabajo automatizados.

Despliegue de Aplicaciones: AWS Lambda para ejecutar código en respuesta a eventos.

Visualización de Datos:

Dashboards: Utilizar servicios como AWS QuickSight o herramientas open-source en instancias EC2, dentro de los límites de la capa gratuita.

Ciclo de Vida del Dato:

Recolección: Automatización y programación de la recolección de datos.

Limpieza y Preparación: Herramientas de procesamiento en Colab para la preparación de datos.

Análisis y Modelado: BigQuery y Colab para el análisis de datos y desarrollo de modelos de ML.

Visualización y Reporte: Creación de dashboards interactivos para presentar KPIs y resultados de análisis.

Monitoreo y Mantenimiento: Servicios de monitoreo en la nube para rastrear rendimiento y costos, ajustando recursos conforme sea necesario.

**OPCION GCP**

Para desarrollar un diseño detallado del proyecto utilizando exclusivamente Google Cloud, se puede seguir la siguiente estructura:

Ingesta de Datos: Utilizar Google Cloud Storage (GCS) para almacenar los datasets en formatos Parquet y JSON. GCS es escalable y ofrece una capa de uso gratuito que es adecuada para el almacenamiento de grandes conjuntos de datos.

Procesamiento y Análisis: Emplear Google Cloud Dataproc para procesar y analizar grandes volúmenes de datos. Dataproc es un servicio de Hadoop y Spark gestionado que permite procesar datos directamente en GCS. Para análisis interactivos y ejecución de queries a gran escala, se puede utilizar Google BigQuery.

Despliegue y Orquestación: Google Cloud Functions y Google Cloud Composer (basado en Apache Airflow) son adecuados para automatizar flujos de trabajo y desplegar aplicaciones. Estos servicios permiten una orquestación eficiente y están integrados con el resto de servicios de Google Cloud.

Visualización de Datos: Google Data Studio o Looker (si está disponible en la capa gratuita) pueden ser utilizados para crear dashboards interactivos que se conecten directamente a BigQuery o a otras fuentes de datos en GCP.

Ciclo de Vida del Dato:

Recolección: Automatización de la ingesta de datos mediante servicios de Google Cloud como Pub/Sub o directamente desde GCS.

Limpieza y Preparación: Herramientas como Google Cloud Dataprep para limpiar y transformar datos antes del análisis.

Análisis y Modelado: Uso de BigQuery ML para modelado directamente en el almacén de datos, y AI Platform para modelos de ML más complejos.

Visualización y Reporte: Data Studio para presentar resultados y KPIs.

Monitoreo y Mantenimiento: Google Cloud Monitoring y Google Cloud Logging para el seguimiento del rendimiento de las aplicaciones y los costos asociados.

**OPCION AWS**

Para diseñar un proyecto detallado utilizando únicamente AWS, siguiendo el ciclo de vida del dato y enfocado en minimizar costos, se propone el siguiente enfoque:

Ingesta de Datos:

Utilizar Amazon S3 para el almacenamiento inicial de datos en formatos Parquet y JSON, aprovechando la capa gratuita que ofrece una solución rentable para el manejo de grandes conjuntos de datos.

Procesamiento y Análisis de Datos:

Emplear AWS Glue para la preparación y carga de datos, lo que permite transformaciones ETL serverless ajustándose a la capa gratuita.

Para el análisis, se puede recurrir a Amazon Athena que permite realizar consultas directamente sobre datos en S3, pagando únicamente por las consultas realizadas.

Despliegue y Orquestación:

Para el despliegue de aplicaciones y modelos, AWS Lambda junto con AWS Step Functions facilita la creación de flujos de trabajo serverless, con un número significativo de ejecuciones gratuitas al mes.

Utilizar Amazon CloudWatch para el monitoreo y mantenimiento, permitiendo ajustar los recursos automáticamente en respuesta a eventos o cambios en la demanda.

Visualización y Presentación de Datos:

Amazon QuickSight proporciona una herramienta de visualización de datos y creación de dashboards intuitiva, con una capa gratuita para análisis básicos.

Para dashboards más complejos o aplicaciones de ML, se puede optar por servicios como AWS SageMaker, que ofrece ciertos beneficios gratuitos para el desarrollo y despliegue de modelos de aprendizaje automático.

Ciclo de Vida del Dato:

Recolección: Configuración de AWS Data Pipeline o AWS Glue para la ingesta automatizada de datos.

Limpieza y Preparación: Uso de AWS Glue DataBrew para preparar y limpiar los datos sin necesidad de escribir código.

Análisis y Modelado: Análisis con Amazon Athena y desarrollo de modelos en SageMaker.

Visualización y Reporte: Creación de dashboards interactivos con QuickSight.

Monitoreo y Mantenimiento: Implementación de Amazon CloudWatch para el monitoreo continuo y la optimización de recursos.

**OTRA OPCIÓN (Quizás más adaptada al modelo)**

* **Arquitectura de Datos:**

Ingesta de Datos:

Crear un mecanismo automatizado utilizando Google Cloud Functions para la ingesta de datos desde las fuentes de Google y Yelp, almacenando los archivos Parquet y JSON en Google Cloud Storage. Esto aprovecha la elasticidad de la nube y la capacidad de procesar grandes volúmenes de datos.

Procesamiento y ETL:

Establecer pipelines de procesamiento de datos en Google Cloud Dataflow para transformar y limpiar los datos. Usar BigQuery para el procesamiento analítico y la preparación de los datos para su análisis y modelado. Este paso garantiza que los datos estén listos para análisis y modelado eficientes.

Almacenamiento y Modelado de Datos:

Utilizar BigQuery para el almacenamiento y modelado de datos, diseñando esquemas de bases de datos optimizados para consultas analíticas y operaciones de ML.

* **Análisis y Procesamiento de Datos:**

Herramientas de Análisis:

Emplear Google Colab para el análisis exploratorio de datos, visualización y modelado estadístico, integrándose con BigQuery y otras herramientas de análisis de Google Cloud.

Modelado de Machine Learning:

Implementar y entrenar modelos de ML en Colab, utilizando TensorFlow o scikit-learn, para luego desplegarlos en Google AI Platform, proporcionando predicciones en tiempo real o por lotes.

* **Visualización y Reporting:**

Dashboards:

Desarrollar dashboards interactivos con Google Data Studio o Looker, conectados a BigQuery para mostrar KPIs y proporcionar insights accionables.

Reportes:

Generar reportes automatizados y dashboards para resumir hallazgos y recomendaciones, distribuidos a través de servicios como Google Pub/Sub o Cloud Functions.

* **Despliegue y Mantenimiento:**

Automatización y Orquestación:

Utilizar Google Cloud Composer para orquestar y automatizar los flujos de trabajo de ETL y ML, asegurando la ejecución regular y la gestión de dependencias.

Monitoreo y Optimización:

Implementar herramientas de monitoreo y logging como Google Operations Suite (antes Stackdriver) para rastrear el rendimiento del sistema y optimizar costos.

* **Ciclo de Vida del Dato:**

*Recolección y Limpieza:*

Ingesta de Datos:

Google Cloud Storage (GCS): Los datos en formatos JSON y Parquet se cargan en GCS, que sirve como el punto de entrada para los datos crudos. Aquí se aplican procesos iniciales de validación para asegurar la integridad de los datos.

Limpieza y Validación:

Google Cloud Dataflow: Se utiliza para aplicar transformaciones iniciales, como la limpieza de datos, eliminación de duplicados, y validación de formatos. Dataflow permite la creación de pipelines de procesamiento de datos escalables y gestionados.

*Procesamiento y Análisis:*

Transformación y Enriquecimiento:

Google Cloud BigQuery: Después de la limpieza, los datos se mueven a BigQuery para su almacenamiento y análisis. Aquí se realizan transformaciones más complejas, agregaciones, y enriquecimiento de datos, aprovechando la capacidad de BigQuery para procesar grandes volúmenes de datos de manera eficiente.

Análisis Exploratorio y Modelado de ML:

Google Colab: Se utilizan notebooks en Colab para el análisis exploratorio de datos (EDA), donde se visualizan y exploran los datos para obtener insights. Además, se desarrollan y entrenan modelos de machine learning utilizando librerías como TensorFlow o scikit-learn.

*Distribución y Reutilización:*

Almacenamiento de Modelos y Resultados:

Google Cloud Storage: Los modelos entrenados y los resultados del análisis se almacenan de nuevo en GCS para su fácil acceso y reutilización.

*Despliegue de Modelos de ML:*

AWS Lambda y Amazon SageMaker: Los modelos de ML se despliegan utilizando AWS Lambda para inferencias en tiempo real y Amazon SageMaker para entrenamientos y despliegues de modelos a mayor escala. Esto facilita la creación de aplicaciones basadas en ML que pueden consumir los modelos directamente.

*Visualización y Reporting:*

Amazon QuickSight y/o Google Data Studio: Para la visualización de los datos y KPIs, se pueden utilizar QuickSight o Data Studio, creando dashboards interactivos que permiten a los usuarios finales explorar los datos y resultados de manera intuitiva.

Ciclo de Retroalimentación y Mejora Continua:

Monitoreo y Optimización: Utilizando Google Stackdriver y Amazon CloudWatch, se establece un proceso de monitoreo continuo para evaluar el rendimiento de los modelos y las aplicaciones, permitiendo la optimización continua de recursos y ajustes en los modelos basados en feedback real.

*Gestión de Proyecto y Documentación:*

Herramientas de Colaboración:

Emplear Google Workspace para la documentación colaborativa y seguimiento del proyecto.

*Control de Versiones y Repositorio:*

Mantener un repositorio en GitHub para el control de versiones, con CI/CD utilizando Google Cloud Build o AWS CodePipeline para la integración y despliegue continuos.

# Equipo de trabajo – roles y responsabilidades

# Cronograma general – GANT

# Análisis preliminar de calidad de datos