Yalbani Aranda

PROYECTO INTEGRADOR

El objetivo de este proyecto e integrar datasets con información relevante a las copas mundiales de futbol, para poder tener una perspectiva más amplia de la cantidad de turistas que se esperan a partir del próximo año.

INGESTA DE DATOS

Las fuentes de datos fueron extraídas de Kaggle. Se usaron datasets reales de datos acerca de asistencia en mundiales de futbol desde 1930. El análisis incluye datos históricos desde 1930, con énfasis en el último mundial en Qatar, proyectando análisis hacia el mundial del 2026. Su carga inicial fue desde una subnet S3 de AWS usando Python y AWS CLI.

Este tipo de almacenamiento es escalable y seguro, S3 protege los datos con cifrado en reposo y en tránsito. Su control de acceso fue mediante IAM y políticas de bucket estándar para obtener la información directamente, y opciones de bloqueo de acceso público para mayor seguridad. Solo usuarios específicos podían modificar los archivos.

Se instaló boto3, ya que se encontraba instalado y habilitado el CLI, para comunicarnos entre S3 y el bucket configurado. Boto 3 decodifica el body o contenido de nuestro objeto JSON, lo recupera y lo convierte en un DataFrame de Pandas para su análisis.

PROCESAMIENTO INTERMEDIO

Se realizó un ETL básico en AWS, que consiste en la limpieza, transformación y agregación de datos para obtener una base más *usable*.

Extracción.

Los datos crudos fueron requeridos directamente de S3 para su procesamiento. Se encuentran en formato .csv.

Transformación.

1. Limpieza de datos.

- ✓ Eliminación de caracteres especiales (y, espacios extras)
- ✓ Estandarización de nombres (estadios, países)
- ✓ Conversión de tipos de datos (strings a numéricos)
- ✓ Normalización de formatos (capacidad, asistencia, budget)

2. Agregaciones.

- ✓ Unificación de datos en 2 bases
- ✓ Renombramiento de columnas para uniformidad
- ✓ Manejo de valores nulos

Carga.

Al finalizar los análisis requeridos, se guardaron las variables finales como DF nuevos, 6 archivos CSV procesados, en una carpeta de datos limpios en S3.

Automatización.

El análisis se dividió en un ETL manual para los datos estáticos históricos, y se implementó una arquitectura Serverless para el análisis de los datos restantes, que son datos que constantemente cambian. Los datasets extraídos de Kaggle con información de estadísticas de jugadores se utilizaron pensando en cambiar esa fuente por la API directa de Transfermarket, que es la base de datos más importante de jugadores y su valor en el mercado, y así tener escalabilidad automática y una automatización completa sin intervención manual.

Se creó una función Lambda utilizando Python, con un Trigger Amazon EventBridge(rate: 1 day), con un timeout de 15 minutos y memoria de 512 MB para que pudiera procesar la gran cantidad de datos de jugadores.

También se limpiaron un poco los datos, ya que, aunque eran más usables y uniformes los datasets, se eliminaron duplicados y sufijos, y se renombraron las columnas para uniformidad. La lambda genera un log JSON con timestamp para registro de actualizaciones, status, resultados y posibles errores. Al final se exportan los datos resultantes a la carpeta de datos limpios, en una carpeta llamada lambda, en formato CSV para compatibilidad con Streamlit.

PROCESAMIENTO DISTRIBUIDO AVANZADO

Se creó un Pipeline en Spark que realizara Joins con datasets complementarios para tener un detalle más preciso acerca de la asistencia a eventos pasados de la copa mundial de futbol.

En Streamlit se utilizó @st.cache_data con una proyección a 10 minutos de actualización automatica, para reducir consultas redundantes a S3 y mejorar tiempo de carga del dashboard, mejorando UX.

VISUALIZACION FINAL

Para que su visualización fuera más accesible, se presenta como un Dashboard interactivo usando Streamlit (framework de Python), con visualizaciones Plotly para gráficos que permiten interpretar los resultados del pipeline y que su lectura sea más interactiva. La publicación del dashboard fue de manera local.

Para levantar el servicio nos comunicamos desde la terminal a AWS, y extraer directamente la información del bucket en S3 para la visualización de los datos finales.

DOCUMENTACION TECNICA

El análisis se basa en un ambiente virtual de Python y una libreta de Jupyter que permite crear kernells o procesadores para ejecutar el código, permitiendo documentar y probar el análisis paso a paso.

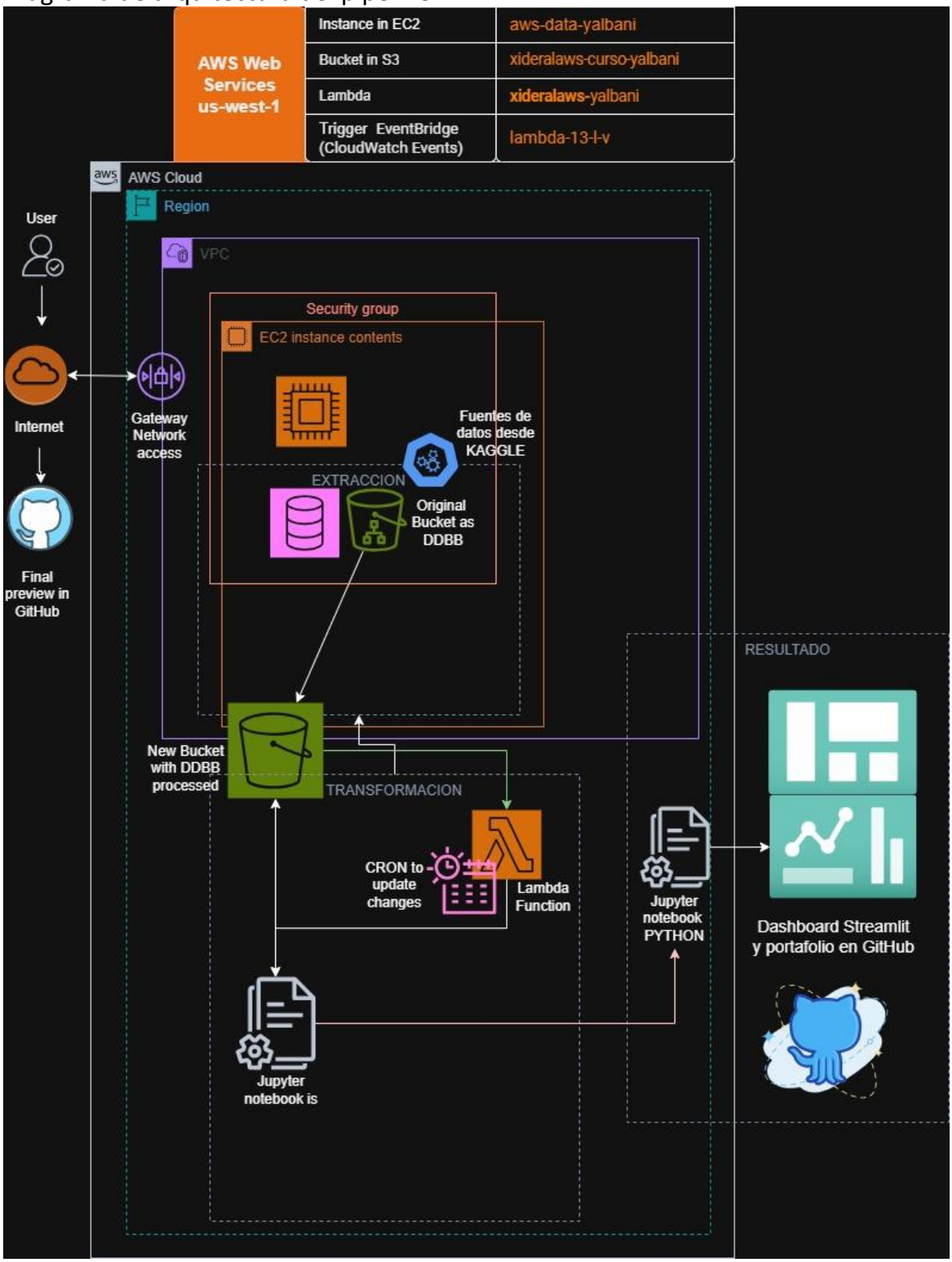
- Networking components

La Región que se utilizó en el proyecto fue us-west-1, del Norte de California.

Se realizó el control de versiones en GitHub, con scripts listos para ser ejecutados de forma reproducible mediante la terminal Ubuntu y AWS CLI, que permite interactuar con AWS, asi como la librería Boto3(acceso via API) de Python. Se usó Amazon VPC como red virtual en AWS Cloud, para mantener conexiones seguras mediante políticas IAM y claves SSH para monitorear el acceso a las instancias. Para acceder a VPN fue de manera privada, que usa encriptación para que los datos viajen de manera segura.

Dentro del VPC (Amazon Virtual Private Cloud) que utilizamos, se utilizaron otro tipo de recursos de AWS. Como *Subnets,* se usó como base de datos EC2 de manera privada, para la información transaccional, y un deploy de manera pública para la visualización del dashboard.

- Diagrama de arquitectura del pipeline.



Este mapa de la red en AWS Cloud, ayuda a entender de manera visual la manera en que la aplicación accede a los servicios, recursos o datos.

Yalbani Aranda

CONCLUSIONES

Los beneficios de esta manera de trabajo y el uso de AWS, es una escalabilidad ilimitada, acceso desde cualquier lugar, respaldos automáticos, procesamiento distribuido, automatización completa y un monitoreo del flujo de trabajo documentado.

Yalbani Aranda

Métricas del Proyecto Finales

Datos Procesados:

- 7 datasets integrados
- 90,000+ registros de jugadores
- 45,000+ partidos históricos
- 40,000+ goles registrados
- 28 eventos mundiales analizados
- 171 estadios catalogados
- 263 países con datos

Análisis Generados:

- 6 análisis principales (manual)
- 3 análisis automatizados (Lambda)
- 1 dashboard con 5 secciones
- 20+ visualizaciones interactivas

Automatización:

- Actualización diaria automática (Lambda)
- Trigger configurado con EventBridge
- Cache inteligente (10 minutos)
- Sin intervención manual requerida

Yalbani Aranda

Insights Clave Obtenidos

1. Asistencia a Estadios:

Promedio histórico de llenado: 95.8%

10 eventos con 100% de capacidad

Mayor asistencia: 173,850 personas (Brasil 1950)

Qatar 2022: promedio 88-94% de llenado

2. Infraestructura Mundial 2026:

Total de estadios disponibles: 171

Capacidad total combinada: ~9 millones de asientos

País con más estadios: Estados Unidos (127)

Estadio más grande: Bristol Motor Speedway (153,000)

Capacidad promedio: 52,631 asientos

3. Análisis de Goles:

Partidos con >4 goles: 5,247 encuentros

Ventaja de local: 70% de victorias en alta anotación

País con mayor ventaja: España (80.5%)

Distribución más común: 5 goles por partido

4. Proyección Financiera Mundial 2026:

Fondo histórico 2022: $1,000 millones USD

Proyección 2026: $1,566.7 millones USD

Crecimiento esperado: 56.67%

Crecimiento histórico total (1982-2022): 4,900%

Tendencia: Crecimiento exponencial sostenido

5. Análisis de Jugadores:

Top 50 G+A concentrado en 15-20 países

Países dominantes: Europa y América del Sur

Goleadores elite (>3 goles): 150+ jugadores

Mayor concentración de talento: Europa Occidental

Yalbani Aranda

RECURSOS:

Fuentes de datos utilizadas:

- Football Stadiums - Football Stadiums.csv

https://www.kaggle.com/datasets/imtkaggleteam/football-stadiums?select=Football+Stadiums.csv

- Qatar 2022 FIFA World Cup Attendance - Attendance Sheet.csv

https://www.kaggle.com/datasets/parasharmanas/qatar-2022-fifa-world-cup-attendance?select=Attendance+Sheet.csv

- FIFA_historical_dataset_latest_2022 - FIFA_history.xlsx

https://www.kaggle.com/datasets/senapatirajesh/fifa-dataset?select=FIFA_history.xlsx

- Football Manager 2023: 90k+ Player Stats - merged_players (1).csv

https://www.kaggle.com/datasets/siddhrajthakor/football-manager-2023-dataset?select=merged_players+%281%29.csv

- International football results from 1872 to 2025 - goalscorers.csv, results.csv

https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017?select=goalscorers.csv

- Ballon d'Or 2025 Nominees Players Standard Stats - ballondor_2025_nominees_dataset.csv

https://www.kaggle.com/datasets/siddhrajthakor/ballon-dor-2025-nominees-players-standard-stats?select=ballondor_2025_nominees_dataset.csv