Proyecto Final

Analítica Predictiva en Entornos de Datos Masivos en la Nube

Objetivo: Diseñar, implementar y documentar una solución de analítica en la nube que trate con datos masivos (≥ 32 millones de registros), integrando componentes de almacenamiento, procesamiento distribuido o incremental, modelado predictivo y despliegue o análisis explicativo, siguiendo las metodologías estudiadas durante el curso.

El proyecto deberá desarrollarse seleccionando una única ruta metodológica entre las dos opciones oficiales:

1. Procesamiento distribuido con PySpark (Dataproc), o
2. Aprendizaje incremental con River y despliegue en Cloud Run.

Ambas rutas son equivalentes en complejidad, esfuerzo y profundidad técnica.

Instrucciones

Cada equipo deberá seleccionar un dataset de gran escala, cumpliendo con **≥ 32 millones de registros**. Se recomienda un dataset que pueda particionarse en múltiples archivos de gran tamaño desde BigQuery. Ejemplos válidos: NYC Taxi, Chicago Taxi, Google Trends, entre otros. El dataset debe exportarse o fragmentarse en Cloud Storage (GCS) para permitir un flujo de procesamiento.

Cada equipo deberá elegir **una ruta** de los siguientes componentes, procurando cubrir una dimensión analítica y una de infraestructura:

Ruta A — Procesamiento Distribuido con PySpark en Dataproc

Esta ruta se centra en la implementación de técnicas de cómputo distribuido para procesar datos masivos utilizando clusters de Dataproc y la API de PySpark.

1. Selección y exportación del dataset hacia GCS
    - Exportación desde BigQuery o desde la fuente de datos seleccionada.
    - Verificación de estructura, consistencia y tamaños.
2. Procesamiento distribuido en Dataproc
    - Creación del cluster.

- Lectura del dataset desde GCS mediante PySpark.
- Aplicación de limpieza, filtrado, transformación y muestreo equivalente al flujo trabajado en clase.

3. Modelado predictivo en PySpark
    - Implementación de un modelo supervisado:
        - *Logistic Regression*, *Linear Regression*, o
        - *One-vs-Rest*, según la naturaleza del dataset.
    - Cálculo de métricas relevantes (Accuracy, Recall, F1 o R^2).
4. Evaluación comparativa entre dos configuraciones de cluster
    - incluir métricas de tiempo y observaciones del Job UI
    - interpretar cómo afecta el tamaño del cluster a la ejecución
    - comentar sobre latencia, paralelismo y escalabilidad
5. Documentación y análisis
    - Interpretación de resultados.
    - Justificación del muestreo.
    - Evaluación del desempeño del modelo.

Evidencias

- Capturas del bucket en GCS.
- Capturas del cluster y del job en Dataproc.
- Script del programa principal de PySpark.
- Tabla comparativa de tiempos y métricas de evaluación.
- Conclusiones.

RUTA B — Aprendizaje Incremental con River y Despliegue en Cloud Run

Esta ruta exige la construcción de un flujo incremental (simulación de streaming) y el despliegue de una aplicación en Cloud Run, conforme a lo desarrollado en la Actividad 2 y las sesiones posteriores.

B.1. Actividades obligatorias

1. Selección y exportación del dataset hacia GCS
    - Mismos requisitos que en la ruta A.
    - El dataset debe estar organizado en múltiples archivos que simulen un flujo de datos entrantes.
2. Implementación de aprendizaje incremental (River)
    - Limpieza y preparación de datos basada en archivos.

   - Entrenamiento incremental archivo por archivo.
   - Evaluación continua mediante métricas como Accuracy, Recall, F1 o R^2.
   - Comparación con un modelo offline equivalente (scikit-learn).
3. Despliegue en Cloud Run
   - Construcción de un microservicio basado en la plantilla ya utilizada en clase.
   - El servicio deberá cargar el modelo incremental o permitir ejecutar predicciones/evaluaciones desde la nube.
   - El repositorio deberá activar CI/CD automáticamente.
4. Documentación y análisis
   - Interpretación del comportamiento incremental.
   - Identificación de drift por mes o partición.
   - Comparación con el modelo batch.

B.2. Evidencias mínimas

- Capturas del bucket en GCS.
- Capturas de la aplicación desplegada en Cloud Run.
- URL funcional del servicio.
- Gráfica de evolución de la métrica incremental.
- Comparación entre el enfoque incremental y el enfoque offline.
- Explicación del código utilizado y conclusiones.

Formato de entrega

- Archivo PDF/Word del reporte: Proyecto_Final_NombreApellido.pdf
- Debe incluir:
  - Dataset utilizado
    - Fuente, tamaño y justificación
  - Descripción de la arquitectura implementada (diagrama)
  - Desarrollo de la ruta elegida
  - QMétricas, gráficas, tablas
  - Análisis crítico: Ventajas y limitaciones del enfoque elegido

Rúbrica

Criterio	Descripción	Puntos
Implementación del pipeline de datos	Ejecución completa y reproducible del flujo técnico (ingesta, preparación, procesamiento y ejecución del pipeline PySpark o River).	40
Modelado predictivo y análisis del desempeño	Correcta implementación del modelo, métricas reportadas, visualizaciones y análisis crítico de resultados o estabilidad.	35
Exposición del trabajo	Claridad y conclusiones bien fundamentadas durante la presentación.	25
	Total	100

Referencias:

- Google LLC (s. f.). *Google Cloud Console*. Recuperado de https://console.cloud.google.com/
- Google Cloud. (2024). *Crea un clúster de Dataproc con la consola de Google Cloud*. Recuperado de https://cloud.google.com/dataproc/docs/quickstarts/create-cluster-console?hl=es-419
- Velázquez O. (2025). *Material de trabajo de la asignatura* [Notebooks]. Google Colab. Recuperado de https://drive.google.com/drive/folders/1jE3cqnPqAXtPCpcZDjMivKE6ipb2v7Wy?usp=sharing