Tarea 3

Bases de Datos a gran escala

mail: claudio.torresf@usm.cl

1. Objetivo

El objetivo de esta tarea es aplicar tus conocimientos en Apache Beam y Apache Airflow para construir una ETL que permita convertir los datos de participación de fans de la Liga de Carreras de Helicópteros (HRL) de Json a Avro, para su posterior uso en el entrenamiento de modelos de Machine Learning.

2. Descripción

La Liga de Carreras de Helicópteros (HRL) recolecta información de la participación de sus fans en sus aplicaciones durante sus transmisiones. El objetivo es entrenar modelos de Machine Learning que permitan predecir si un fan utilizará el servicio de predicciones y/o comprará artículos de Merchandising con el fin de ofrecer banners de publicidad personalizados.

Los datos se recolectan en formato Json, con la siguiente estructura:

- FanID: identificador único del fan.
- RaceID: identificador de la carrera visualizada por el fan.
- Timestamp: momento en el que el fan inicio la transmisión de la carrera, en formato '%Y-%m-%d %H:%M:%S'.
- ViewerLocationCountry: país desde donde visualiza la carrera.
- DeviceType: tipo de dispositivo desde el cual visualiza la carrera.
- EngagementMetric_secondswatched: cantidad de segundos que un fan ha visualizado la carrera.
- PredictionClicked: True si accedió a la sección de predicciones, False en caso contrario.
- MerchandisingClicked: True si accedió a la sección de compra de Merchandising, False en caso contrario.

Los ingenieros de ML le han pedido convertir la columna Timestamp a Unix timestamp en milisegundos, por lo que se ha agregado una nueva columna: Timestamp_unix.

La plataforma de ML soporta datos en formato Avro, los datos de salida deben utilizar el siguiente esquema Avro:

```
{
  "type": "record",
  "name": "FanEngagement",
  "fields": [
    {"name": "FanID", "type": "string"},
    {"name": "RaceID", "type": "string"},
    {"name": "Timestamp", "type": "string"},
    {"name": "Timestamp_unix", "type":{
           "type": "long",
           "logicalType": "timestamp-millis"
         }},
    {"name": "ViewerLocationCountry", "type": "string"},
    {"name": "DeviceType", "type": "string"},
    {"name": "EngagementMetric_secondswatched", "type": "int"},
    {"name": "PredictionClicked", "type": "boolean"},
    {"name": "MerchandisingClicked", "type": "boolean"}
  ]
}
```

Se incluye un fichero JsonL con datos de prueba.

3. Actividades

Para los requerimientos mencionados, deberás (100 puntos):

1. Construir una ETL en Apache Beam que realice la conversión de los archivos Json a Avro con el esquema entregado. Su ETL deberá calcular el campo Timestamp_unix. Debe tener en cuenta que los datos de entrada y de salida pueden estar en un bucket y/o en el sistema de ficheros local. (40 puntos)

2. Construir un Dag en Airflow que permita ejecutar en forma diaria el procesamiento de los archivos Json de entrada. (40 puntos) Al finalizar el job de Apache Beam deberá notificar a un tópico de Kafka que lo datos están disponibles para consumo, con el siguiente mensaje Json:

   ```
   {
       "event_type": "data_processing_completed",
       "data_entity": "FanEngagement",
       "status": "success",
       "location": path_or_bucket,
       "processed_at": processed_timestamp,
       "source_system": pipeline_name
   }
   ```

   Con

- **path_or_bucket:** path o URI de los datos procesados.
- **processed_timestamp:** momento en el que se terminó el procesamiento de los datos y se envió el mensaje, en formato '%Y-%m-%d %H:%M:%S'.
- **pipeline_name:** nombre de su Dag.

3. Construya un workspace usando Visual Studio Code y devcontainers en el cual se pueda ejecutar su código en local, incluyendo los servicios necesarios para su correcto funcionamiento, debe agregar un archivo Readme.md con las indicaciones para ejecutar su Dag. (10 puntos)

4. **Conclusiones:** desarrolle un pequeño informe con los principales desafios a los que se enfrentó para resolver esta actividad. (10 puntos)

4. Consideraciones

- La tarea puede ser desarrollada en grupo de máximo 3 estudiantes.
- El informe debe poseer el nombre de todos los integrantes del equipo de trabajo. Este debe ser en formato PDF.
- Debe entregar su código fuente.
- Se debe enviar un informe por equipo en la plataforma AULA.
- Está permitido utilizar Inteligencias Artificiales para el desarrollo de la tarea, pero debe agregar el prompt utilizado en el informe.

Fecha entrega: 29 Junio 2025