Proyecto

Bases de Datos a gran escala

mail: claudio.torresf@usm.cl

1. Objetivo

El objetivo de ese proyecto es construir una infraestructura y una estrategia integral que permita gestionar, procesar, analizar y compartir de manera eficiente la enorme cantidad de datos de la empresa TerramEarth para generar valor de negocio significativo.

Este proyecto esta basado en el caso de estudio TerramEarth del examen Google Cloud Professional Cloud Architect (`https://services.google.com/fh/files/blogs/master_case_study_terramearth.pdf`).

2. Descripción

TerramEarth fabrica maquinaria pesada para las industrias minera y agrícola. Actualmente cuenta con más de 500 concesionarios y centros de servicio en 100 países. Su misión es desarrollar productos que aumenten la productividad de sus clientes.

Actualmente hay 2 millones de vehículos TerramEarth en funcionamiento, y observamos un crecimiento anual del 20 %. Los vehículos recopilan datos de telemetría de numerosos sensores durante su funcionamiento. Un pequeño subconjunto de datos críticos se transmite desde los vehículos en tiempo real para facilitar la gestión de la flota. El resto de los datos de los sensores se recopila, se comprime y se carga diariamente cuando los vehículos regresan a su base. Cada vehículo suele generar entre 200 y 500 megabytes de datos al día.

2.1. Datos

Para el presente proyecto analizaremos los datos de las mquinarias agricolas con el fin de optimizar uso de fertilizantes y la mejora del rendimiento de los cultivos. Además de analizar la duración y fallas comunes de las maquinarias, con el fin de evaluar fallas de diseño y mejorar los problemas más comunes.

2.1.1. Datos críticos

Los datos críticos corresponden a los datos operacionales de cada maquinaria, es decir, el estado de cada sensor del vehiculo. Adicionalmente, se puede agregar información de diagnostico y errores. Esta información se envia cuando una maquina se enciende, apaga o se activa un testigo de error.

Estos son los datos recopilados durante el funcionamiento de la máquina, que proporcionan información sobre su rendimiento y actividad:

- Vehicle ID: Identificador único de la máquina (p. ej., AGR-001, MIN-005).

- Timestamp: Hora exacta en que se registró el punto de datos (p. ej., 2025-05-25T14:30:15Z).
- GPS Coordinates: Latitud y longitud de la ubicación de la máquina (p. ej., Lat: -33.4489, Lon: -70.6693).
- Engine RPM: Revoluciones por minuto del motor (p. ej., 1850 RPM).
- Engine Load: Porcentaje de la capacidad del motor utilizada (p. ej., 75
- Fuel Level: Porcentaje de combustible restante (p. ej., 68
- Fuel Consumption Rate: Litros por hora o galones por hora (p. ej., 12,5 L/h).
- Speed: Velocidad de avance de la máquina (p. ej., 8,2 km/h).
- Odometer Reading: Distancia total recorrida (p. ej., 15 432,7 km).
- Operating Hours: Total de horas que la máquina ha estado en funcionamiento (p. ej., 2100,5 h).
- Hydraulic Pressure: Presión en el sistema hidráulico (p. ej., 2500 psi).
- Hydraulic Fluid Temperature: Temperatura del fluido hidráulico (p. ej., 85 °C).
- Engine Coolant Temperature: Temperatura del refrigerante del motor (p. ej., 90 °C).
- Battery Voltage: Voltaje de la batería de la máquina (p. ej., 24,1 V).
- Gear Engaged: Marcha actual de la transmisión (p. ej., 3.ª, Avance).
- Accelerator Pedal Position: Porcentaje de presión del pedal (p. ej., 60
- Brake Pedal Position: Porcentaje de presión del pedal (p. ej., 0

Los datos de Diagnóstico y Error son cruciales para el mantenimiento predictivo y la identificación de posibles problemas antes de que provoquen fallos catastróficos. Los datos capturados son:

- Error Code: Código de error específico de la máquina (p. ej., P0420, Advertencia de Sobrecalentamiento del Motor).
- Warning Light Status: Encendido/Apagado para varias luces de advertencia del tablero (p. ej., Luz de Comprobación del Motor Encendido, Presión de Aceite Baja Apagado).
- Component Temperature: Temperatura de un componente especifico, como un rodamiento, transmisión, etc. (p. ej., Temperatura del Rodamiento Delantero Izquierdo: 70 °C).
- Vibration Data: Lecturas sin procesar del sensor de vibración o un índice de vibración calculado (p. ej., Vibración_X: 0,5 G, Vibración_Y: 0,3 G).

- Component Wear Indicators: lecturas de sensores diseñados para medir el desgaste de piezas como pastillas de freno, profundidad de la banda de rodadura de los neumáticos, etc. (por ejemplo, BrakePad_FrontRight_Thickness: 8,5 mm).

2.1.2. Datos de sensores

Además de la telemetría principal del vehículo (o datos críticos), la maquinaria agrícola envía datos de los sensores de sus implementos acoplados (p. ej., sembradoras, pulverizadoras, cosechadoras):

- Implement ID: Identificador único del implemento acoplado (p. ej., PLNTR-001, SPRAYR-003).
- Section Control Status: Activado/Desactivado para secciones individuales de la barra de pulverización o hileras de sembradoras (p. ej., Section1 On, Section2 Off).
- Flow Rate (Fertilizer/Pesticide): Litros por minuto o galones por minuto (p. ej., 5,3 L/min).
- Application Rate (Fertilizer/Pesticide): Litros por hectárea o galones por acre (p. ej., 150 L/ha).
- Seed Rate: Semillas por metro cuadrado o semillas por acre (p. ej., 7,5 semillas/m^2).
- Yield Monitor Data: Datos del Monitor de Rendimiento (Cosechadoras)
  - Yield Moisture Content: Porcentaje de humedad del cultivo cosechado (p. ej., 18,2
  - Yield Mass Flow: Kilogramos por segundo o bushels por minuto (p. ej., 2,1 kg/s).
  - Yield Rate Per Hectare: Rendimiento estimado en kg/ha o bushels/acre (p. ej., 9800 kg/ha).
- Soil Sensor Data: Datos del Sensor de Suelo (si está integrado):
  - Soil Moisture Percentage: Porcentaje de humedad del suelo (p. ej., 35
  - Soil Temperature Celsius: Temperatura del suelo (p. ej., 22,5 °C).
  - Soil EC: Conductividad eléctrica del suelo (p. ej., 1,2 dS/m).
- Soil Nutrient Levels: niveles de nutrientes (nitrógeno, fósforo y posatio) del suelo (p. ej., nitrógeno 15 ppm).

2.1.3. Ejemplo


```
{
  "Vehicle_ID": "AGR-001",
  "Timestamp": "2025-05-25T14:35:22Z",
  "GPS_Coordinates": {
```

```
  "Latitude": -33.4501,
  "Longitude": -70.6705
},
"Engine_RPM": 1920,
"Engine_Load_Percentage": 80,
"Fuel_Level_Percentage": 67,
"Fuel_Consumption_Rate_L_hr": 13.1,
"Speed_km_h": 7.8,
"Odometer_km": 15433.2,
"Operating_Hours": 2100.6,
"Hydraulic_Pressure_psi": 2600,
"Hydraulic_Fluid_Temp_C": 87,
"Engine_Coolant_Temp_C": 91,
"Battery_Voltage_V": 24.0,
"Gear_Engaged": "3rd",
"Accelerator_Pedal_Position_Percentage": 65,
"Brake_Pedal_Position_Percentage": 0,
"Diagnostic_Data": {
  "Error_Code": null,
  "Warning_Lights": {
    "CheckEngineLight": "Off",
    "LowOilPressure": "Off"
  },
  "Component Temperatures_C": {
    "Engine": 92,
    "Transmission": 85,
    "HydraulicSystem": 88
  },
  "Vibration": {
    "Vibration_Index_X": 0.48,
    "Vibration_Index_Y": 0.52,
    "Vibration_Index_Z": 0.55
  }


},
"Implement":{
  "Implement_ID": "PLNTR-001",
  "Section_Control_Status": {
    "Section1": "On",
    "Section2": "On",
    "Section3": "Off"
  },
  "Flow_Rate_L_min": 45.0,
  "Seed_Rate_seeds_per_sq_m": 7.6,
```

```
    "Yield_Monitor_Data": {
      "Yield_Moisture_Content_Percentage": 12.5,
      "Yield_Mass_Flow_Rate_kg_s": 2.3,
      "Yield_Rate_Per_Hectare_kg_ha": 1500.0
    },
    "Soil_Sensor_Data": {
      "Soil_Moisture_Percentage": 18.0,
      "Soil_Temperature_C": 20.5,
      "Soil_Nutrient_Levels": {
        "Nitrogen_ppm": 25,
        "Phosphorus_ppm": 15,
        "Potassium_ppm": 30
      }
    }
  }
}
```

3. Actividades

Debes realizar las siguientes actividades (100 puntos):

1. **Diseñe** una arquitectura de datos para procesar y almacenar la información de TerramEarth. Aspectos a considerar: (30 puntos)

   - Sistemas fuente (source) que generan los datos a ingestar. Definir que tipo de serialización será usada para recibir los datos.
   - Ingesta de los datos: describa los factores de diseño que debe considerar. ¿Qué forma de ingestión de datos se utilizará?
   - Transformación.
   - Servir los datos. ¿Cómo serán entregados los datos a los usuarios finales?
   - Seguridad: ¿Comó se protegerán los datos en tránsito?

   Debe incluir un diagrama que muestre la arquitectura propuesta.

2. **Defina** las tecnologias y servicios que se deben utilizar para implementar su arquitectura. (30 puntos)

3. **Desarrolle** un POC utilizando Python en el cual implemente una ETL/ELT para procesar un conjunto de datos de prueba. Se incluye un script Python con un generador the datos aleatorios en base al esquema presentado. (30 puntos)

4. **Conclusiones.** (10 puntos)

Finalmente, deberás elaborar un informe conciso y bien estructurado. Además, debes presentar tu solución en una presentación de no más de 15 minutos.

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.
- 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 informe y presentación: 05 de Julio de 2025