**Challenge Avanzado: Análisis de Grandes Bases de Datos con PySpark, Relacionado a la Tesis**

1. **Introducción:**

Este challenge avanzado permitirá a los estudiantes aplicar PySpark para analizar grandes volúmenes de datos relacionados con un tema cercano a su investigación de tesis, integrando los conocimientos adquiridos en los challenges anteriores.

1. **Objetivo general:**

Realizar un análisis avanzado utilizando PySpark, relacionando el proyecto con su tema de tesis.

1. **Objetivo específico:**

Utilizar APIs avanzadas como Spark Streaming o Spark ML para trabajar con una gran base de datos en tiempo real o construir un modelo predictivo complejo. Los estudiantes deberán crear una presentación de su análisis y presentarla al grupo.

1. **Base de datos:**

Base de datos: Dependiendo del tema de tesis de cada estudiante. Sugerencias:

* [Data del Banco Mundial](https://databank.worldbank.org/)
* Base de datos de Clima

1. **API y herramientas:**

Spark Streaming para análisis en tiempo real o SparkML para modelos avanzados de predicción.

1. **Presentación:**

Crear una presentación en PowerPoint o similar y presentarla frente al grupo.

1. **Repositorio GitHub:**

Subir el código, la presentación y la documentación a un repositorio en GitHub.

1. **Métricas de evaluación:**

* Relación clara con el tema de tesis (20%)
* Uso avanzado de PySpark (25%)
* Análisis detallado y conclusiones (25%)
* Calidad de la presentación frente al grupo (20%)
* Organización y limpieza del código en GitHub (10%)