## ACTIVIDAD SESIÓN PROCESAMIENTO DE DATOS

El Ministerio de Educación ha publicado un informe sobre las carreras universitarias con mayor cantidad de estudiantes inscritos en Chile en los últimos años. Se ha recopilado un conjunto de datos que incluye información sobre la carrera, la universidad, la cantidad de inscritos y el área de conocimiento.

Tu tarea es procesar esta información utilizando **Spark SQL y DataFrames** para responder preguntas clave y generar un informe procesado.

**Dataset proporcionado: carreras.json**

Formato de los datos:


```
[
 {"id": 1, "carrera": "Ingeniería Comercial", "universidad": "U. de Chile", "inscritos": 3200, "area": "Económicas y Administrativas"},
 {"id": 2, "carrera": "Derecho", "universidad": "PUC", "inscritos": 2900, "area": "Ciencias Sociales"},
 {"id": 3, "carrera": "Medicina", "universidad": "U. de Concepción", "inscritos": 2500, "area": "Salud"},
 {"id": 4, "carrera": "Psicología", "universidad": "U. de Santiago", "inscritos": 1800, "area": "Ciencias Sociales"},
 {"id": 5, "carrera": "Ingeniería Civil", "universidad": "UTFSM", "inscritos": 3100, "area": "Ingeniería y Tecnología"}
]
```


**INSTRUCCIONES**

**1.- Creación de la sesión Spark (1 punto)**

- Crea una sesión de Spark con el nombre "AnalisisCarreras".

**2.- Carga de datos en un DataFrame (1 punto)**

- Carga los datos desde el archivo carreras.json en un DataFrame de Spark.
- Muestra los primeros registros.

**3.- Exploración del DataFrame (1 punto)**

- Imprime el esquema del DataFrame para visualizar los tipos de datos de cada columna.

**4.- Consultas con Spark SQL (3 puntos)**

Realiza las siguientes consultas usando Spark SQL:

a) Obtener todas las carreras con más de 2500 inscritos.

b) Contar cuántas carreras pertenecen a cada área de conocimiento.

c) Mostrar las universidades que ofrecen más de una carrera en la lista de datos.

**5.- Creación de una Función Definida por el Usuario (UDF) (2 puntos)**

- Crea una UDF que clasifique las carreras según la cantidad de inscritos:
    - "Alta demanda" si tiene más de 3000 inscritos.
    - "Media demanda" si tiene entre 2000 y 3000 inscritos.
    - "Baja demanda" si tiene menos de 2000 inscritos.
- Aplica la UDF al DataFrame y agrega la nueva columna "demanda".
- Muestra el DataFrame actualizado.

**6.- Guardado de datos en Parquet (2 punto)**

- Guarda el DataFrame transformado en un archivo Parquet llamado "carreras_procesadas.parquet".

**INSTRUCCIONES ADICIONALES:**

- Puntos totales = 10 puntos.
- Comprimir el archivo en formato .zip o .rar.
- Subir el archivo a la plataforma.