Examen Tercera evaluación

Supongamos que trabajamos con una empresa dedicada a la investigación y análisis de datos relacionados con Pokémon. Esta empresa recopila datos de diversas fuentes, como archivos JSON, archivos de texto, MongoDB. El objetivo es realizar análisis de datos para comprender mejor el mundo de Pokémon, incluido el comportamiento de los entrenadores, las tendencias de los Pokémon más populares y la distribución geográfica de las especies.

EXTRACCIÓN de datos

Será necesario generar los siguiente archivos de datos que tendréis que cargar posteriormente en S3.

Archivos JSON

Utilizaremos archivos JSON que contienen información detallada sobre diferentes especies de Pokémon, incluidas sus estadísticas, habilidades y tipos.
Nombre del archivo: pokemon_data.json
Atributos:

- Nombre (string)
- Tipo (string)
- Estadísticas (objeto JSON con atributos numéricos como HP, Ataque, Defensa, etc.)
- Habilidades (array de strings)
- Evoluciones (array de objetos JSON con el nombre y tipo de Pokémon evolucionado)

Ejemplo:

```
{
  "Nombre": "Bulbasaur",
  "Tipo": "Planta/Veneno",
  "Estadísticas": {
    "HP": 45,
    "Ataque": 49,
    "Defensa": 49,
    "Velocidad": 45
  },
  "Habilidades": ["Espesura", "Clorofila"],
  "Evoluciones": [
    {"Nombre": "Ivysaur", "Tipo": "Planta/Veneno"},
    {"Nombre": "Venusaur", "Tipo": "Planta/Veneno"}
  ]
}
```

Archivos de texto

Además de los archivos JSON, utilizaremos archivos de texto que contienen registros de batallas Pokémon, incluyendo información como el nombre del evento jugado, el nombre del entrenador, los Pokémon en su equipo y el resultado de la batalla.

Nombre del archivo: battle_records.txt

Atributos:

- Nombre del entrenador (string)
- Equipo de Pokémon (array de strings con nombres de Pokémon)
- Resultado de la batalla (string)
- Evento (string)

Ejemplo:

```
Evento: Torneo Pokémon Ciudad Pallet
Nombre del entrenador: Ash Ketchum
Equipo de Pokémon: [Pikachu, Charizard, Bulbasaur, Squirtle,
Jigglypuff]
Resultado de la batalla: Ganó
```

MongoDB

La base de datos MongoDB contiene datos sobre eventos especiales de Pokémon, como torneos en línea, distribuciones de Pokémon raros y eventos de la serie de televisión.

Base de datos: pokemon_events_db

Colección: events_collection

Atributos:

- Evento (string)
- Fecha (string o tipo de fecha)
- Descripción (string)

Ejemplo:

```
{
  "Evento": "Torneo Pokémon Ciudad Pallet",
  "Fecha": "2023-05-15",
  "Descripción": "Gran torneo Pokémon celebrado en Ciudad Pallet"
}
```

Apache Kafka

Para mejorar el procesamiento en tiempo real, utilizaremos Apache Kafka para recibir eventos de batallas Pokémon en vivo. Cada mensaje representará una batalla con información sobre el entrenador, el equipo utilizado y el resultado.

Topic de Kafka: pokemon_battle_events

Estructura del mensaje:

- Timestamp (string en formato ISO 8601)
- Evento (string)
- Nombre del entrenador (string)
- Equipo de Pokémon (array de strings)
- Resultado de la batalla (string)

Ejemplo:

```
{
  "Timestamp": "2025-04-03T14:23:00Z",
  "Evento": "Torneo Pokémon Ciudad Verde",
  "Nombre del entrenador": "Gary Oak",
  "Equipo de Pokémon": ["Blastoise", "Arcanine", "Alakazam",
"Gengar", "Rhydon"],
  "Resultado de la batalla": "Perdió"
}
```

TRANSFORMACIÓN de datos

En esta etapa, los datos ya estarán en s3 con LocalStack y tendréis que prepararlos para el apartado de análisis,**tendréis que eliminar los elementos repetidos.**

LOAD: Data Warehouse

Data loading

Los datos transformados se cargarán en el Data Warehouse, hacedlo con localstack o postgres, para su análisis posterior en 3 tablas distintas que responderán a las preguntas del Data analytics. Solo poner la información de cada tabla que sea interesante para resolver estas preguntas.

Data analytics

Usando Apache Spark tenéis que obtener los datos a través de Data Warehouse y realizar consultas que contengan análisis avanzados sobre los datos almacenados en el almacén de datos.

a. Características de Pokémon:

- ¿Cuáles son los Pokémon con mayor HP?
- ¿Qué Pokémon tiene el mayor ataque?
- ¿Cuáles son las habilidades más comunes entre los Pokémon?

b. Comportamiento del entrenador:

- ¿Quién es el entrenador con más victorias registradas?
- ¿Cuál es el equipo de Pokémon más utilizado por los entrenadores ganadores?
- ¿Existe alguna correlación entre el tipo de Pokémon y el resultado de la batalla?

c. Eventos especiales de Pokémon:

- ¿Cuántos eventos especiales se han registrado en la base de datos?
- ¿Cuál fue el evento más reciente?
- ¿Qué descripción tiene el evento con más batallas?

Estructura del proyecto

Para este proyecto tendréis que seguir la siguiente estructura:

data_bda/: Esta carpeta contiene todos los datos necesarios para el análisis.

- **json/**: Aquí se almacena el archivo pokemon_data.json.
- **text/:** Aquí se almacena el archivo battle_records.txt
- **mongodb/:** Esta carpeta almacena el archivo pokemon_events.json.

data_generation/: Esta carpeta contiene todos los archivos python que generen los archivos anteriores, **tienen que generarlos en la carpeta especificada.**

apps_bda/: Esta carpeta contiene el resto de archivos py.

- **data_integration.py**: Almacena datos en s3
- **data_transformation.py**: eliminar repetidos, organizar datos.
- **data_load.py**: Guarda los datos en las 3 tablas distintas de Postgres/Redshift.
- **data_analysis.py**: documento donde se realiza el análisis de datos.

Si existe algún archivo más que sea necesario **comentarlo** con el profesor.

Evaluación

Me tendréis que hacer un zip con todos los archivos utilizados para hacer el examen.