

Práctica COVID

Vamos a implementar un proyecto real, para ello se requiere tener un repositorio git donde se almacene todos el código y el power BI sin datos, es decir solo se subirá el modelo y para cargar datos deberá tener un parámetro que cargue los datos.

Debemos generar 2 ficheros independientes uno que contenga una tabla de ciudad con ID_ciudad, id, coordenadas y otra con país con id_pais, id, las tablas tienen que tener coherencia, es decir España puede ir con Madrid, valencia pero no con París. Poner 6 ciudades

Para la api vamos a usar Py para atacar a la api y obtener datos de una Fecha concreta (recordad que ya no se actualiza por tanto tenéis que mirar desde que fecha hasta que fecha hay datos), ese fichero tiene que ser guardado en json.

Una vez el fichero descargado habrá que leer el json con spark y agregar las columnas id_pais, id_ciudad y Factor multiplicador, este factor va a multiplicar todos los valores numéricos, la idea es tener para el mismo día datos de los 6 ciudades y guardarlo en parquet.

El siguiente paso utilizar pentaho o un script spark para leer ese parquet para añadir F_procesado (Cuando se procesa) y las columnas aaaa, mm, dd de la fecha de los datos. Y guardarlo en un parquet histórico (podéis tener 1 para cada país-ciudad)

Tener datos para 4 meses

Para la parte de PBI tendrá que leer los ficheros parquet históricos, y los ficheros de ciudad y país, donde vais a montar el modelo de datos, aplicar un RLS con 4 roles (1 rol que vea 2 países, 1 que lea todo y los 2 últimos que solo vean un país). se tiene que guardar sin datos, es decir tiene que tener un parámetro en la carga de datos que al ser igual cargue datos y si no que lo deje vacío como esto 0=1 (0 datos) 1=1 (carga todos los datos) y generar un cuadro de mando informativo con lo que consideréis oportuno.

Consejos:

1. Crear Repositorio GIT
2. API-File-Json -> API.COVIDTRACKING.COM leer un poco la api para entender que se quiere y como se tiene que hacer.
3. Json – parquet
 - a. Usar Spark para leer y agregar país, ciudad, factor multiplicador anonimo
4. Script/ETL leer Parquet/json y añadimos las columnas Fecha_procesado, aaaa, mm, dd y hacer un append para tener el histórico.
5. Cualquier fichero procesado tiene que moverse a una carpeta con nombre old o procesado para saber que fichero se ha procesado.
6. Podéis usar una base de datos para cargar el histórico