Tabla de contenido

Instrucciones de uso

Requisitos:

- Tener instalado Python 3.x en su máquina
- Tener instalado Spark 2.x en su máquina

Sobre el programa

Este es un programa cuyo objetivo es crear un archivo *.parquet* dada una muestra y el schema de esta muestra. El programa está conformado por dos archivos *.py*: ***paths.py*** y ***precess.py***. Es importante que estos archivos permanezcan en un mismo directorio.

Preparando los archivos

El programa está diseñado para trabajar con dos archivos *.txt*, estos archivos deben seguir una estructura determinada para que puedan ser procesados de forma correcta.

La muestra

El primer archivo debe ser la muestra con la cual se desea construir el archivo *.parquet*. A continuación se muestra una imagen con el formato adecuado para este archivo:

```
|contact_center_type|operation_currency_type|gl_fixed_asset_account_id|load_date |
|9999               |                       |1 11                     |2019-12-31|
|9999               |                       |1 1101                   |2019-12-31|
|9999               |                       |1 1101 01                |2019-12-31|
|9999               |                       |1 1101 02                |2019-12-31|
```

Es importante que este archivo cumpla con esta estructura y que sea guardado como *.txt*.

El schema

El segundo archivo debe ser el schema de la muestra de la cual se desea construir el archivo *.parquet*. A continuación se muestra una imagen con el formato adecuado para este archivo:

```
contact_center_type: string
operation_currency_type: string
gl_fixed_asset_account_id: string
load_date: date
```

Es importante que este archivo cumpla con esta estructura y que sea guardado como *.txt*.

Estableciendo los parámetros

Se debe abrir el archivo ***paths.py*** con su editor de texto favorito y editar los siguientes parámetros:

- **sample:** es la dirección donde se encuentra el archivo <tu_muestra.txt>.
- **schema:** es la dirección donde se encuentra el archivo <tu_schema.txt>.
- **partition:** es la partición por la cual se particionarán los datos. Es importante que esta sea un campo que pertenece a la muestra.
- **output:** es la dirección donde se guardará su archivo parquet.

Luego de establecer los parámetros de acuerdo a sus necesidades se deberán guardar los cambios realizados en el archivo.

Ejecutando el programa

Una vez establecidos los parámetros solo resta ejecutar el archivo process.py para que cree el archivo .parquet.

¿Cómo ejecutar el archivo process.py?

Abra una consola o terminal según aplique a su SO y diríjase hasta la dirección donde se encuentra el archivo **process.py**. Una vez allí ejecute el siguiente comando:

```
〉 python process.py
```

Si ha realizado todos los pasos anteriores correctamente el programa creará el archivo *.parquet* de forma exitosa.

Es importante que tenga en cuenta que todos los archivos que estén en la ruta establecida en **output** serán borrados debido a que el modo de escritura es overwrite. Se recomienda crear un directorio nuevo para sus salidas.

Contacto

Para reportar cualquier fallo o realizar recomendaciones por favor contactarme.

Nombre: José Miguel Moya
Correo: josemiguel.moya.curbelo.contractor@bbva.com