**ETL INSIVUMEH**

Descarga, tratamiento de control de calidad y carga de los datos climàticos diarios de las estaciones convencionales.

{Se debe incluir cada cuanto se ejecutará}

Horario de ejecución

7: 00 - 7:40 am con intervalos de 5 minutos.

13:00, 13:30 p.m

18:00 - 18:30 p.m

19:00 p.m

{Parámetros de ejecución: Como archivos de fechas, conexiones, carpetas de entrada, salida, kettle.properties, etc}

* Token de Kobotoolbox.
* API de Google Drive activada para Google Sheet.
* Secret\_cliente brindado por la GCP
* Sistema Linux.
* Carpeta compartida en Google Drive.

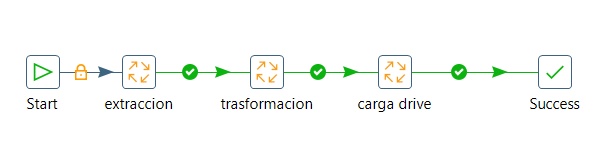
{Repositorio de código en GIT}

{Como se ejecuta}

Se ejecuta a través de línea de comando con Crontab con el sistema operativo Linux.

**Componentes del proceso**

**ETL INSIVUMEH**



***Job: Job\_***

Este proceso es el encargado de ejecutar cada nodo del ETL, es decir, el job\_extraccion, job\_transformacion y job\_carga\_drive para generar datos climáticos con un control de calidad robusto.

1. **Extracción de datos**

Imagen que contiene Escala de tiempo

Descripción generada automáticamente

***Job: Job\_extraccion***

Este proceso es el encargado de descargar los datos de Kobotoolbox a través de la API:

**1.1** Star inicia el proceso.

**1.2** La transformación Shell con alias “extraccion” se encarga de ejecutar el script en Python que descargar los datos generados a través de Kobotoolbox empleando el API respectiva.

**1.3** Success.

1. **Transformación de datos**

Imagen que contiene Gráfico de dispersión

Descripción generada automáticamente

***Job: Job\_transformacion***

**2.1** Star inicia el proceso.

**2.2** La transformación Shell con alias “t1\_orden” ejecutar el script en Python que descarga los datos de Kobotoolbox a través de la API y autorización con el Token. Empleando un diccionario para cambiar el nombre de la estación con el ID interno. Elimina y ordena para dejar solamente las columnas necesarias.

**2.3** La transformación Shell con alias “t2\_salida” ejecutar el script en Python que unifica los datos de los formularios y saca el promedio de los tres horarios según la variable.

**2.4** La transformación Shell con alias “t3\_dir\_viento” ejecutar el script en Python que toma los valores de dirección de viento y genera el dato medido de la respectiva variable.

**2.5** La transformación Shell con alias “t4\_anulacion” ejecutar el script en Python que basandose en los ingresos del formulario Anulación transforma el dato a un valor nulo. Esto con su respectiva razón de cambio.

**2.6** La transformación Shell con alias “t5\_salida\_general” ejecutar el script en Python que genera el .CSV sin las columnas extras.

**2.7** La transformación Shell con alias “t6\_correccion\_lluvia” ejecutar el script en Python que toma la fecha del ingreso del dato y le resta un 24 horas, dado que la lluvia corresponde al día anterior. Genera el .CSV que utilizará en el siguiente paso.

**2.8** La transformación Shell con alias “t7\_validacion\_rangos” ejecutar el script en Python que compara los valores ingresados con el límite mínimo y el máximo que se encuentran en la Masterdata. Generando un documento llamado “Fuera de rango” y el que contiene los datos preliminares.

**2.9** La transformación Shell con alias “t8\_orden\_columnas” ejecutar el script en Python que toma el nombre de las columnas, según este lo ordena bajo el formato establecido. También, realiza un renombre de las variables.

**2.10** La transformación Shell con alias “t9\_avisos telegram” ejecutar el script en Python que se encarga de enviar por mensaje en Telegram a través de un bot, alertas tempranas sobre el envío de datos duplicados, fuera de rango, estación faltante, entre otras.

**2.11** Success.

1. **Carga de datos**

Imagen que contiene Diagrama

Descripción generada automáticamente

***Job: Job\_carga\_drive***

Este proceso es el encargado de cargar el documento de documento\_drive y Fuera de rango a una carpeta compartida en el Google Drive:

3.1 Star inicia el proceso.

3.2 La transformación Shell con alias “carga\_drive” se encarga de cargar los documentos que contienen los datos diarios y fuera de rango a una carpeta compartida a los horarios respectivos.

3.3 Success.