**Prueba de desarrollo para VinkOS**

Se tiene el requerimiento de crear un proceso ETL que integre la información de visitas de un sitio web contenida en archivos planos.

Estos archivos se encuentran en el servidor remoto (8.8.8.8) y se tiene acceso por sftp, contamos con las credenciales y el directorio donde se encuentran es /home/vinkOS/archivosVisitas

Los archivos tienen la extensión txt y la estructura del nombre es “report\_”+consecutivo+.”txt”, no se tiene el número exacto de archivos que se generan por día.

El proceso ETL deberá hacer lo siguiente:

* Ir todos los días al directorio para buscar los archivos
* Validar el layout de los archivos
* Validar la información a cargar:
  + Email correcto
  + Fechas en formato dd/mm/yyyy HH:mm
* Cargar la información en 3 tablas mysql

visitante

* + - email, fechaPrimeraVisita, fechaUltimaVisita, visitasTotales, visitasAnioActual, visitasMesActual
  + estadística
    - email,jyv,Badmail,Baja,Fecha envío,Fecha open,Opens,Opens virales,Fecha click,Clicks,Clicks virales,Links,IPs,Navegadores,Plataformas
  + errores: registros con error
* Borrar los archivos cargados en el origen
* Hacer un backup de los archivos cargados en un zip y mandarlo al directorio local /home/etl/visitas/bckp
* Restricciones:
  + No se debe cargar un archivo más de una vez
  + El proceso es responsable de la administración de los archivos (borrado en origen y backup en destino)
  + El servidor donde se ejecuta el proceso ETL se utilizará como storage del backup
  + En la tabla visitante, solo hay un registro por email, si no existe se agrega, si existe se actualizan los valores de fechaUltimaVisita, visitasTotales, visitasAnioActual, visitasMesActual.
    - fechaUltimaVisita: la fecha en formato yyyymmdd de la última visita
    - visitasTotales: Conteo del número de visitas desde su primera visita.
    - visitasAnioActual: Conteo del número de visitas del año actual.
    - visitasMesActual: Conteo del número de visitas del mes actual.
* Inclusión de tecnologías Big Data, describir cómo cambiaría el flujo del proceso si el destino fueran tablas en Impala o Hive en un cluster Hadoop.

El objetivo del test, es detallar el flujo del proceso del ETL a construir, es decir, dónde habrá puntos de control, qué validaciones se utilizarán, cómo se administrarán y notificarán los errores, cómo y en qué orden se llenarán las tablas destino, si hay reglas, cuáles serán y dónde se aplicarán.

Adicionalmente, se debe describir el proceso para:

1. Los elementos que se deben considerar para la liberación del proceso a un ambiente productivo.
2. La operación del ETL ya que en este ejemplo, quien lo desarrolla no opera, es decir, qué herramientas, métricas ó elementos se deben facilitar al equipo operativo para tenga visibilidad de la correcta ejecución del proceso y en caso de un incidente, saber que acciones realizar.

Los archivos de ejemplo se encuentran en la carpeta denominada “txt”.