Automatización de la Parametrización de Fuentes para la Ingesta al Lago de Datos de Davivienda

Jesus Alberto Blanco Barbosa

Facultad De Ingeniería De Telecomunicaciones, Universidad Santo Tomas

Dc. Juliana Arévalo Herrera

Bogotá 2021

Agradecimientos

El proceso de elaboración de este trabajo fue de mucha paciencia por parte de Juliana Arévalo Herrera, directora de mi monografía por esto quiero darle agradecimientos especiales por ayudarme a mejorar y aprender en la creación de este documento.

Dedicatoria

Este trabajo se lo dedico a Dios y a mi familia por ese afecto y amor que me brindan incondicionalmente en los momentos difíciles de mi carrera universitaria, en especial a mi hermana que sin su apoyo no hubiera sido posible.

Tabla de Contenido

Índice de Figuras y Tablas

Tablas

Figuras

Anexos

Resumen

Se realiza un análisis de los pasos que se tienen en el flujo de ingesta de datos al lago de Davivienda, en la que se ve que la parametrización de fuentes que se viene utilizando en el lago de datos es un proceso muy manual y propenso a errores esto genera aumentos de los tiempos que se tiene fijados para cada una de las etapas, del procesamiento de datos. En este contexto, se plantea el desarrollo de una aplicación para parametrizado automático, que reduzca los errores que se estaban presentando y la parametrización manual.

Abstract

An analysis of the steps in the flow of data ingestion to the Davivienda lake is carried out, in which it is seen that the parameterization of sources that has been used in the data lake is a very manual and error-prone process. this generates increases in the times set for each of the stages of data processing. In this contexthe development of a new parameterization process that is more efficient, autonomous and that reduces the number of errors that are occurring is proposed.

1. Introducción

En este documento se presenta el proyecto que se trabajó en el banco Davivienda durante las prácticas empresariales, Davivienda es una entidad financiera, sus productos se enfocan en atender las necesidades de las personas, empresas, sector minero y energías limpias con innovación constante. Durante las prácticas empresariales se trabajó en el área de Tecnología Planeación y Riesgo, un área que tiene a su cargo varias etapas del proceso de ingesta de información al lago de datos, en las practicas se trabajó en las siguientes etapas de entendimiento, refinamiento, prototipado y documentación prototipado. El proyecto del cual se habla en este documento se realizó en la etapa de prototipado, con el fin de tener una mejora en tiempos y los errores que se estaban presentando.

Se inicia con una contextualización de todos los temas que serán utilizados durante el desarrollo del documento, se continua con un análisis dentro del área para identificar posibles mejoras en los pasos que se tienen en la actualidad en el proceso de la ingesta al lago de datos, con el objetivo de proponer una mejora del proceso que se tiene actualmente para la ingesta de los datos. Se concluye que uno de los pasos**,** específicamente el de parametrizado genera una gran pérdida de tiempo y errores.

Durante el proceso se evidencia que el proceso de parametrización que se viene adelantando se realiza muy manual, aumentando la posibilidad de error y generando pérdidas en tiempo por lo tanto se generan pérdidas económicas.

Frente a este escenario, se plantea el desarrollo de un proceso que facilitara a que la parametrización que se venía utilizando quede obsoleta, esto porque la nueva forma de parametrización supera todas las problemáticas que se venían presentando de manera recurrente al momento de parametrizar una fuente por el método tradicional.

Se espera que la automatización del parametrizado de fuentes reducirá de manera sustancial los tiempos que se gastaba una persona en parametrizar una fuente de muchos campos y mejorará la calidad pues la automatización toma los datos directos del diccionario lo que evitará los errores humanos que se presentaban durante el proceso.

2. Objetivos

2.1 General

Automatizar la parametrización de fuentes en el aprovisionamiento de ingesta a el lago de datos de Davivienda, para el proceso de prototipado de fuentes, que permita reducir el tiempo requerido en al menos un 25%, agilizando el proceso de parametrización de fuentes.

2.2 Específicos

- Presentar los conceptos de bases de datos, lago de datos y las herramientas usadas en el trabajo.
- Identificar el proceso y las etapas actuales que se tienen en la ingesta al lago de datos explicando las partes que son implicadas y los tiempos requeridos
- Proponer el nuevo proceso de parametrización de fuentes, para la ingesta de datos en el banco.
- Elaborar un manual que facilite el entendimiento del nuevo proceso de parametrización de fuentes en la ingesta a el lago de datos.

3. Capítulo 1: Conceptos Básicos para la Ingesta de Datos

Los datos en la actualidad son el activo más importante que tienen las empresas, conocer la importancia de los datos ayudará a tener un mejor entendimiento de los temas y del porqué (Yalcin et al., 2022a). Los datos en las empresas están desde el inicio de la misma, para tener un mayor control y una centralización de los datos se crean las bases de datos**.**

Una base de datos es un conjunto de datos que se almacenaron sistemáticamente con una estructura y contexto, para ser utilizados posteriormente por la empresa en las actividades que se requieran**,** las bases de datos tienen dos vertientes bases de datos relacionales y las no relacionales.

Las bases de datos relacionales son aquellas que tienen una organización de las información que se está almacenando, y proporciona a las personas los datos que están relacionado entre sí, como su nombre lo indica se basan en un modelo relacional, cada fila de una tabla relacional tiene un registro y una ID único, las columnas contiene los atributos asignados por los datos, cada uno de estos registros tiene un valor para los atributos de las columnas, por esto se les llama bases de datos relacionales(Samydurai et al., 2022).

Las bases de datos no relacionales son un sistema de almacenamiento que no pose una estructura definida como lo son las relacionales, las bases de datos no relacionales no tienen un registro por fila y no se va a contar con atributos por campos, como si lo tienen las bases de datos relaciones, en pocas palabras las bases de datos no relacionales se emplean cuando no se satisface un modelo de relación. (Samydurai et al., 2022).

En la actualidad, todas las organizaciones emplean especialistas en análisis de datos, con el fin de desarrollar herramientas que ayuden a la toma de decisiones en las empresas, la ayuda

consiste en minimizar la incertidumbre en la toma de una decisión, identificar qué quieren los usuarios, oportunidades de negocios, minimizar pérdidas y realizar mejoras en los procesos (Bayrak, 2021).

La recolección y análisis de datos cambió la forma en la que se tomaban las decisiones en las empresas. Este cambio ayudó a que muchas empresas se mantengan innovando constantemente sus servicios y productos, así no quedaran obsoletas por el mercado tan cambiante que se tiene hoy(Müller et al., 2018).

Por esta razón la mayoría de las empresas en la actualidad, almacenan toda información de sus clientes en bases de datos, para después realizar procesos de analítica con el fin de encontrar pepitas de oro en la información que se está almacenando, como lo mencionan **(**Pigni et al., 2016**).** Todos tienen un afán por extraer pepitas de oro de esas montañas de datos que tienen almacenadas.

Sin duda los datos en la actualidad son uno de los activos más importante de las empresas, ya que, si las empresas no implementan una analítica de datos estarán en desventaja con respecto a otras empresas que estén en su mismo nicho de mercado que si utilicen analítica de datos.

Davivienda que emplea una cultura de empresa innovadora, se sumó a la recolección y análisis de datos hace varios años atrás, por esto creo un lago datos o como se conoce en el banco DTLK. En éste, se almacena toda la información que llega a el banco, con el fin de emplear un análisis a los datos posteriormente, de esta manera mejorar los servicios que se prestan, un ejemplo exitoso fue la aplicación de Daviplata, la cual inicio dentro del banco y se mejoró con la información del lago de datos, una vez que se terminaron y se aprobaron todas las pruebas, se da luz verde para que la aplicación salga al público. Hoy en día esta aplicación cuenta con más de 5

millones de usuarios, estos usuarios aportan datos que se analizan para las mejoras de la aplicación Daviplata así estará siempre a la vanguardia de lo que quieren sus usuarios.

Davivienda cuenta con su propio lago de datos, se le llama lago de datos al repositorio en el cual se almacenan todos los datos que llegan a Davivienda, estos datos son diversos y provienen de diferentes orígenes, estos llegan primero a la zona cruda del lago de datos, donde están los datos sin procesar, una vez se realiza un curado de la información los datos pasan a la otra sección del lago llamada zona de curado y por último llegan a la zona de consumo, esta es la zona donde ya los datos son útiles para diferentes áreas del banco, así cada área de Davivienda se mantiene innovando y buscando darle lo mejor a sus clientes.

En la imagen 1 se puede ver los orígenes, los medios de trasmisión usados y las diferentes zonas del lago de datos. La primera es donde interactúa las personas que realizan el prototipado, llamada zona cruda de la información, en esta zona debe llegar la información sin alteraciones, por parte de las etapas por las que pasa antes de llegar a la zona cruda, como sale la información del origen debe llegar a la zona cruda. La siguiente zona es la de curado, como su nombre lo indica en esta zona es donde se realizan las alteraciones y las modificaciones de los datos, para dejarlos en la última zona que es la de consumo, cuando ya los datos se encuentran en la zona de consumo es en esta zona donde las áreas interesadas toman los datos y dan uso en generación de gráficas, modelos analíticos que cimenta la toma de decisiones, nuevas ideas de negocio, mejoras en los procesos y muchos casos de uso más.

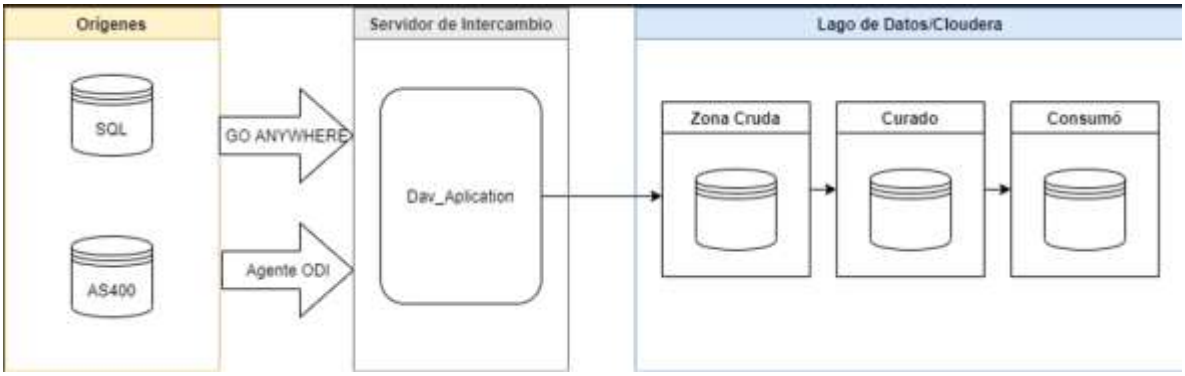


Imagen 1: Arquitectura del flujo el lago de datos

El Origen de la información son diferentes bases de datos o aplicativos con los que cuenta Davivienda, algunos de estos orígenes son: AS400, SQL-Net, SCACS entre otras más. El origen cómo su nombre lo indica es de donde sale la información con la que se alimenta el lago de datos, esa información es trasportada por medio unos caminos Go Aniwehere o Agente de ODI. En la etapa de integración y desarrollo IT ellos deben garantizar que el camino que se va tomar este funcional, si es el caso contrario ellos deben levantar la conexión entre el origen y el servidor de intercambio.

La información llega al servidor de intercambio y se almacena en una carpeta llamada *Landing*, en esta carpeta se tiene un archivo muestra de la fuente que se va prototipar, este archivo es utilizado realizar las pruebas de cargue y comprobar que la parametrización de la fuente es concordante con lo que se tiene en el archivo muestra. una vez se termina las pruebas se realiza el cargue de ese prototipo, ese primer cargue se debe realizar de manera manual ejecutando el ejecutando el Dav_Aplication.

El Dav_Aplication es una herramienta desarrollada por el banco encargada de monitorear y ubicar las fuentes que se encuentra en la carpeta *landing*, para realizar el cargue de la información a el lago de datos del banco. En prototipado se garantiza con el primer cargue que la

fuente este bien parametrizada, para que una vez quede autónomo el cargue de la fuente no genere errores. Su objetivo es la disposición de las fuentes en la zona cruda del lago de datos, para posteriormente realizar modelos analíticos. La interfaz que facilita la interacción es HUE, todas las zonas del lago se pueden visualizar con HUE si se tiene permisos de vista, la imagen 2 es una captura de la interfaz.

La interfaz HUE no fue desarrollada por el banco, se desarrolló por otra empresa llamada Cloudera. Esta empresa es desarrolladora de software que permita el análisis de grandes cantidades de datos, HUE es uno de sus desarrollos y este permite la interacción con las diferentes bases de datos que se tiene en el banco. Gran parte de los desarrollos que ha realizado la empresa Cloudera son de código abierto, lo que permite una integración con otras tecnologías que están enfocadas en la analítica de datos, desarrolladas por ellos u otras empresas, el que sea de código abierto permite moldearse y mejorase, a las preferencias del cliente, es este caso Davivienda.

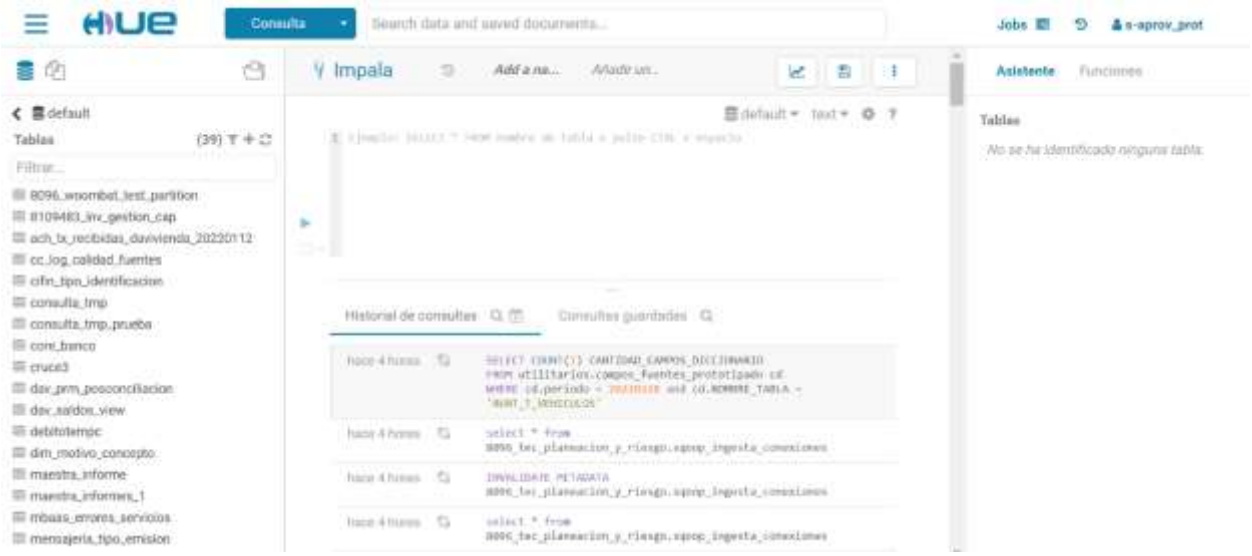


Imagen 2: Interface HUE

4. Capítulo 2: Etapas en la Ingesta al Lago Datos

Todo el flujo que sigue una fuente para la ingesta al lago de datos se evidencia en la imagen 3, cada una de las fuentes que ingresan al flujo de ingesta, debe aprobar cada una de las etapas que se muestras en la imagen 3, cada etapa tiene un color que hace referencia al equipo que interviene, alguna de las etapas se desarrolla en conjunto por varios equipos.

Que se llame flujo no quiere decir que una fuente no pueda ser devuelta a etapas a las cuales ya fue aprobada, por el contrario, estos casos son muy comunes, debido a cambios de información o errores que se cometieron en una etapa anterior, lo que se quiere con el nuevo proceso de parametrización es evitar que las fuentes no vuelvan a la etapa de prototipado por errores de parametrización.

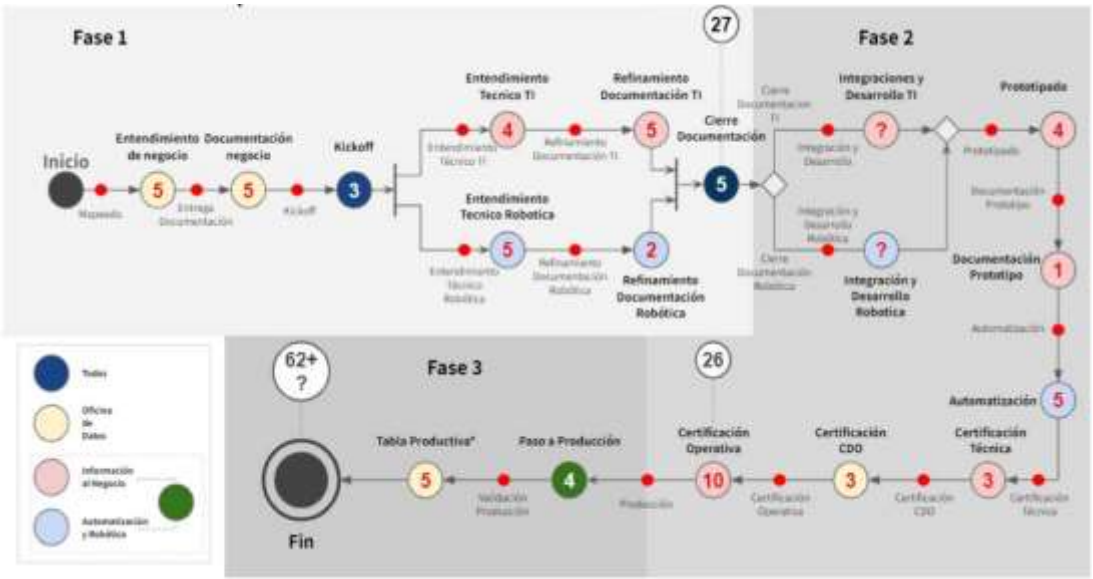


Imagen 3: Flujo de ingesta/elaborado por Davivienda

El tiempo que puede demorar una fuente en cada una de las etapas esta establecidos en los Acuerdos de Niveles de Servicio (ANS), estos son los tiempos que se acordaron para cada una de las etapas en la ingesta al lago de datos. El infligir este tiempo puede representar llamados de atención por parte de jefe encargado de la ingesta de datos.

Estos tiempos se cuentan en días, el número de días para cada etapa se visualiza en la imagen 3 en los círculos debajo del nombre de cada etapa. Los ANS ayudan a tener un estimado de cuantos días faltan para que una fuente quede productiva en el lago de datos, si una fuente empieza a tener demoras en una etapa en la cual ya tiene el ANS vencido, el banco puede recurrir a llamados de atención al líder del área.

El proceso tiene una duración total de 62 días, ese es el tiempo óptimo para que una fuente esté disponible en el área de consumo del lago de datos y pueda ser utilizada.

En la práctica este tiempo aumenta un poco, ya que se pueden presentar retrasas en diferentes actividades tales como, no tener una conexión habilitada con el origen de la información y el encargado del origen demore en responder, que se comentan errores en el parametrizado de la fuente, que el diccionario presente campos con novedades, entre otros posibles errores.

De todas las etapas de la imagen 3, para la cual se realizó el desarrollo y se estará ejecutando el nuevo proceso de parametrización, es para la etapa de prototipado. En esta etapa se están incurriendo en el incumplimiento de los ANS en 1 o más días. Gracias al análisis de las fuentes que volvía a la etapa de prototipado, se evidencia que volvía por errores en la parametrización de fuentes. El nuevo proceso de parametrización llega como una solución a esta problemática que se está presentando.

En prototipado se lleva a cabo la elaboración de ese primer modelo de la fuente que se va ingestar a el lago de datos, para la elaboración de ese primer diseño se deben realizar los siguientes pasos:

1) Parametrizado, en este paso agregamos cada uno de los campos que tiene la fuente, el tipo de dato, su longitud, el nombre de ese campo en el lago. Con el Parametrizado

estamos formado el esqueleto que va tener la tabla en el lago de datos, por eso es uno de los pasos más importantes, ya que, si ese esqueleto queda mal, lo que pasara es que la fuente no se cargara de manera correcta a el lago de datos y genera errores e imposibilitara su consumo.

2) El siguiente paso es ejecutar un script (Shell), lo que se quiere es evaluar la información campo a campo en busca de caracteres especiales (Signos, símbolos, figuras entre otros muchos más) que puedan generar un error, en el cargue a el lago de datos, el script lo identificara y los cambiara por uno que no presente un error, se deja claro que el script se crea para cada fuente y ninguna es igual. Mas adelanten en el texto explicamos que es un script con un lenguaje más de desarrollo.

4.1 Proceso de Creación y Parametrizado de las Fuentes

Para un mejor entendimiento de la problemática que se estaba presentado, se debe presentar el método actual que se tiene para la disponibilidad de una fuente en la zona cruda del lago de datos.

Todo inicia con el líder de la etapa de prototipado, es el encargado de centralizar todas las fuentes que llegaron a prototipado y realizar la distribución de las fuentes para cada uno de los integrantes que participan en la etapa de prototipado. Mayormente las fuentes son entregadas por la etapa anterior los días lunes y miércoles, pero pueden llegar otros días de la semana, el líder de la etapa debe llevar el control.

Todos los días de la semana, en las horas de la mañana se realiza una reunión con el fin de llevar un control de los avances realizados, la reunión inicia preguntando qué fuentes se prototiparon el día anterior, cuales se trabajarán ese día y qué dificultades se presentaron, si

llegaron fuentes nuevas se realiza la una nueva distribución de estas de una manera equitativa con el fin de no sobrecargar a nadie.

Una vez el líder envía el nombre de las fuentes que le pertenecen a cada integrante de la etapa de prototipado, la persona inicia a prototipar la fuente, debe realizar los pasos de imagen 4.

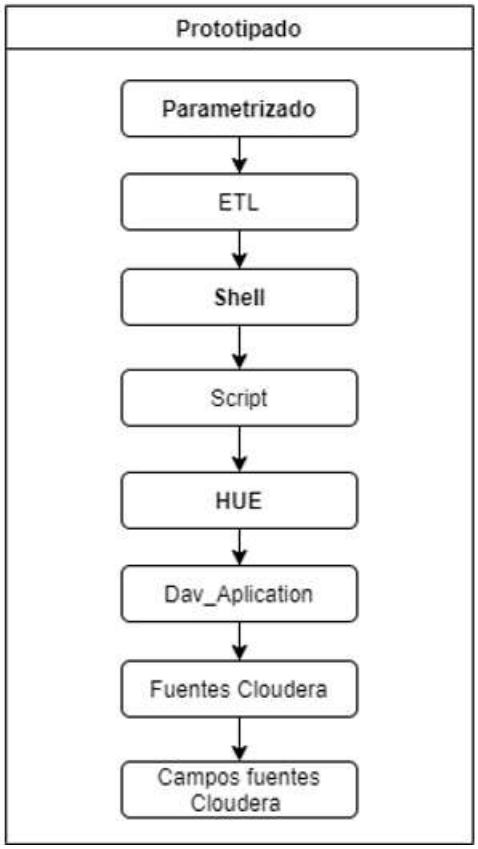


Imagen 4: Fases de prototipado/elaborado propia

Parametrizado de una fuente, es el primer paso como se observa en la imagen 4. Se cuenta con la interfaz de usuario ETL_CLOUDERA, la cual desarrollo el banco para facilitar el proceso de parametrización de fuentes nuevas o para editar las ya creadas, son múltiples pestañas por esta razón se elabora un diagrama imagen 5.

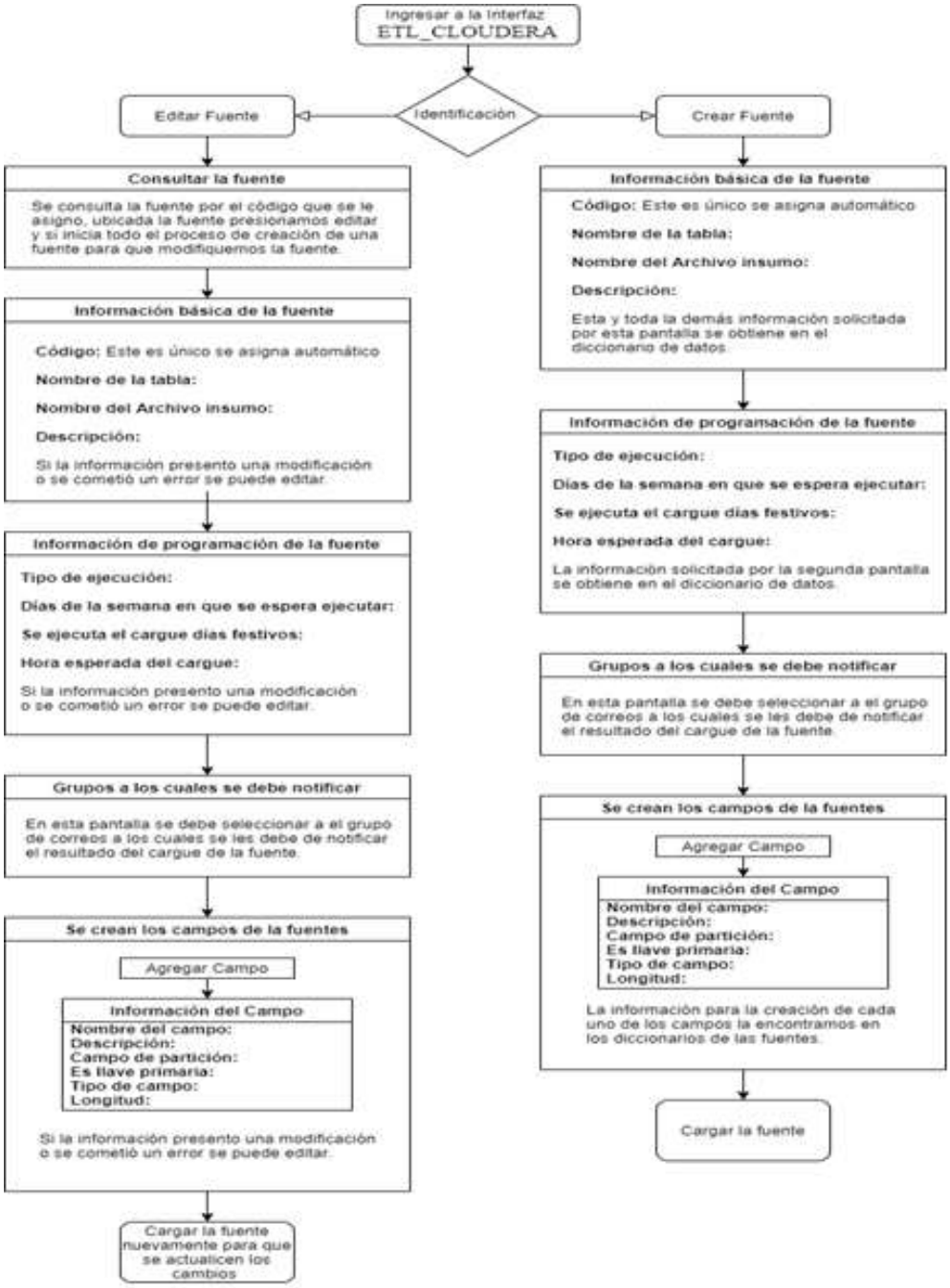


Imagen 5: ETL_CLOUDERA/Elaboración propia

Cada fuente que ingresa a el lago de datos debe pasar por todos los pasos que se observan en la imagen 5 ETL _Cloudera, todos estos pasos los realiza una persona de forma manual, el tiempo que se puede demorar una persona en parametrizar una fuente depende de lo extensa que

esta sea, un promedio es de entre 30 minutos y 4 días. Una vez finalizado el proceso de parametrización, se continua con la creación del script de limpieza de la fuente.

Existen varios tipos de script como las de línea de texto, graficas o de lengua natural, la que se utiliza en el banco es una script de line de texto, esta es una interfaz muy manual en la que se escriben las ordenes que se quieren ejecutar para la fuente que se está trabajando, se tiene un estándar por lo que no se debe elaborar todo el script.

Se debe realizar una adaptación específica para cada fuente que se está prototipando, ejemplos de las ordenes que se pueden dar son, cambiar los espacios por punto y coma, quitar las comas, quitar comillas, quitar el signo pesos, cambiar la letra Ñ por N, entre otras, la limpieza que se va a realizar depende de que caracteres especiales trae la fuente consigo, la limpieza se debe realizar para cada fuente. Lo que se quiere con la limpieza es quitar todos los caracteres especiales que trae la fuente y que muchas veces pueden generar error al momento de cagar la información al lago de datos. Para modificar la script abre con un editor de código (Notepad++), se agrega el nombre de la fuente que se está trabajando y las ordenes que queremos ejecutar y guardamos**,** ejemplos de partes de una script imagen 6.

```
        echo "EJECUTANDO LIMPIEZA "$NOMBRE_ARCHIVO >> $LOG
./$SHELL_LIMPIEZA $NOMBRE_ARCHIVO
RESULTADO_PROCESO=$?
if [ $RESULTADO_PROCESO -ne 0 ];
then
    echo "FALLO SHELL DE LIMPIEZA. ERROR   : " $RESULTADO_PROCESO >> $LOG
    exit 1
fi
            gsub(";"," ",$33); #IDTERMINAL
```

Imagen 6: Ejemplo de una parte de un Script

En la parte de gsub lo que se está realizando es que se eliminando él punto y como, la coma, el espacio y decimales, una vez realizado esa limpieza se debe realizar el cargue de la fuente a la zona cruda del lago de datos.

Terminada el script de limpieza, la fuente queda lista para que una persona se dirija a la interfaz de HUE imagen 7 y disparemos el Dav_Aplication para cargar una de las dos tablas: fuentes_cloudera: Esta tabla contiene información básica de la tabla que se desea crear en él lago de datos, como el nombre, particionamiento, aplicativo al que pertenece, periodicidad de cargue entre otros. El diccionario es el encargado de abastecer esta información que es necesaria para el primer paso de la disponibilidad de la información en la zona cruda del lago de datos.

Campos_fuentes_Cloudera: Esta tabla contiene el detalle de los campos que pertenecen a las tablas, como nombres, tipos, longitudes de los campos, transformaciones, entre otros.

Una vez se realice el cargue de la fuente, la información quedara disponible en la zona cruda del lago de datos.

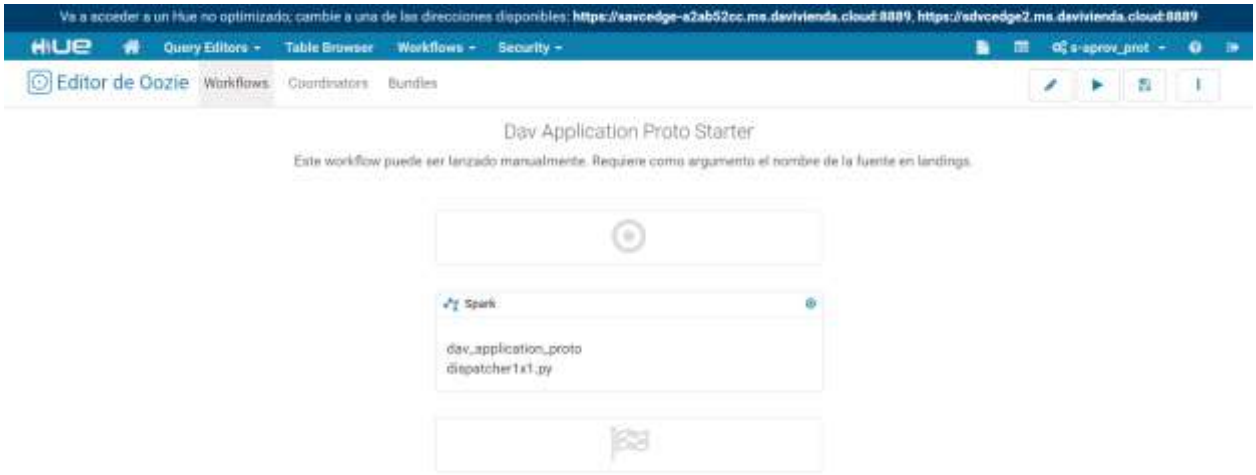


Imagen 7: Dav_Application

El proceso de edición de una fuente es muy similar, la diferencia radica en la primera pantalla del ETL_Cloudera se elige la opción editar fuente, la siguiente pantalla pedirá ingresar el código de la fuente, una vez se ingresa el código único de la fuente, el ETL_Cloudera dejara visualizar las ventanas como si se estuviera creando una fuente nueva como se puede ver en la imagen 5, solo que esta vez ya se tiene información, se edita el campo errado se guarda y se realiza de nuevo a él cargue de la tabla al lago de datos para que la información se actualice.

Tanto en el proceso de modificación como el proceso de creación de una tabla, en la actualidad se están realizando de manera manual, se parametriza la tabla siguiendo los requerimientos del diccionario, El parametrizado de las fuentes por medio del ETL_Cloudera es muy manual, visto de manera rápida no tiene complejidad alguna, pero no es lo complejo del proceso, si no, lo extenso que este se puede volver por ejemplo una fuente de 500 campos se debe parametrizar uno a uno cada campo, lo que aumenta considerablemente la posibilidad de cometer errores, en el parametrizado.

Con este proceso se tiene una relación con el número de campos con respecto al número de errores y el tiempo, si el número de campos es mayor, aumenta el tiempo y el porcentaje de error, y si el número de campos es menor, es menor el tiempo y el porcentaje de error. Todo esto factores se relaciona al factor humano, ya que si el numero de campos es pequeño el cansancio es poco, pero a medida que aumenta el numero de campos el cansancio en las personas en mayor y tienden a cometer mas errores.

Errores Comunes

Errores comunes en el prototipado de las fuentes y afectaciones.

Errores	**Afectación**
Cambiar nombre destino	Cuando el recurso sea utilizado se encontrará la novedad, lo que ocasionará que la fuente vuelva para ser ajustada
Espacios en el nombre destino	Se genera error al momento de cargar los datos y no se podrá consultar la tabla en HUE
Cambiar el tipo de campo	Se puede generar un error de perdida de información, o el otro caso es que quede la fuente disponible, pero el formato del campo no es adecuado, el usuario genera un incidente para arreglar el formato del campo
No especificar la longitud del campo	Perdida de información, error porque el tamaño de la longitud no es el apropiado, reingreso a prototipado.
Cambiar el valor de la longitud	Reingreso a prototipado para realizar la modificación y perdida de la información.

Nota. **Fuente:** elaboración propia

Todas las afectaciones van a generar un impacto negativo en tiempos ya que se debe buscar una solución a la afectación, por lo general todas las afectaciones se terminando en un reingreso de la fuente a la etapa de prototipado, aumentando tiempos y recursos.

5. Capítulo 3: Presentación del Seguimiento al Nuevo Proceso de Parametrización

Debido a la creciente demanda de prototipado de fuentes y que cada vez estas presentan muchos más campos por parametrizar, las devoluciones de las fuentes de las etapas posteriores a la de prototipado estaban en aumento y se estaban incumpliendo los ANS de las fuentes. Debido a esta problemática creciente se plantea un proyecto que ayude en la reducción de tiempos y que disminuya significativamente los errores comunes en la parametrización.

Se establecen unas reuniones de trabajo en las cuales se quería dar una claridad de lo que se quería y en que parte del proceso de prototipado era que se estaban generando mas reprocesos, se identifica que el proceso de parametrización era donde más se estaba fallando. Esto se identifica por las causticas de las tablas que era devueltas a el proceso de prototipado, de igual forma de estas reuniones se concluye, quienes estarían al frente del proyecto para su desarrollara y ejecución.

Se asignan a dos personas y si era requerido se contaban con 50 horas, por la duración del proyecto con el proveedor Asesoftware, para consultas y ayudas en desarrollo, una vez se tenía el equipo de trabajo y se tenía claridad que proceso que se quería mejorar, se pasó a realizar una ruta de trabajo en la que se visualizara el avance del proyecto y el tiempo que este duraría. Se utilizo una plantilla que tiene el banco para dar seguimiento a los proyectos imagen 8, algo muy importante es que el proyecto se dividió en dos fases, esto para tener una parte funcional del nuevo proceso de parametrización lo antes posible.

La primera fase esta 100% funcional y es la que se expone en este documento, con la novedad que esta tiene una parte manual, la segunda fase del proyecto es quitar esa parte manual de la primera fase, y apagar totalmente la anterior forma de parametrizado la cual era por medio

ETL_Cloudera, con el fin de que todo quede automatizado y se ejecute a una hora específica del día y no depende de que una persona.

Esto no le resta valor a la primera parte del proyecto, pues esta fue muy positiva desde su salida sean parametrizado alrededor de 300 fuentes, con este nuevo proceso y ninguna de estas ha presentado alguna de las casuísticas comunes de errores, la ruta de trabajo es la de imagen 8.

Imagen 8: Presentación de seguimiento

Se establecen actividades generales, donde cada uno de estas tiene un tiempo y porcentaje de avance, todo este porcentaje suma en el porcentaje total, en la parte inferior se cuenta con 3 recuadros, lo que se quiere es exponer las dificultadas, logros y próximos pasos. La presentación con la que se llevó el avance del proyecto es la imagen 8.

Para dar seguimiento a los proyectos del área, se planeó cada lunes en las horas de la mañana una reunión, liderada por German Gustavo jefe del área, en la reunión se presentaba la imagen 8 y se exponían los avances que se realizaban y las dificultades que se habían presentado. Así el

conocía el estado de cada proyecto, y las dificultades que este estaba presentando o había presentado, ayudando cuando era requerido a mitigar esas dificultades si estaba a su alcance.

Como se puede visualizar en la imagen 8 y como se mencionó anteriormente el proyecto ya termino en su primera parte y actualmente se encuentra en uso y los resultados que este a dado son muy positivos con alrededor de 300 fuentes parametrizadas y errores comunes en cero en las 300 fuentes que se han parametrizado, donde varios de estas fuentes tienen más de 100 campos. Gracias al buen impacto, el jefe del área motivo a la innovación en los procesos que se desarrollan en el área, con el fin de reducir los reprocesos y estar en constante crecimiento.

El nuevo proceso en ejecución demora entre 5 a 20 minutos y se está realizando una vez por semana, se cargan alrededor de 30 fuentes por semana, donde muchas de estas tienen más 100 campos.

Eficiencia del 98.96% (según se muestra en la siguiente sección) con respecto a el proceso que se estaba utilizando anteriormente, pues antes cada persona que prototipaba una fuente realizaba el parametrizado de la fuente y esto generaba pérdidas de tiempo, errores, ineficiencia en el proceso parametrizado, esto por el factor humano, hoy con el nuevo método solo es un persona la que está encargada de parametrizar todas la fuentes que llegan durante la semana y el tiempo en que demora en parametrizar todas las fuentes es mucho menor a lo que demoraba una persona en parametrizar solo una fuente con el método anterior.

6. Capítulo 4: Manual y Resultados de la Implantación

Para facilitar la transferencia del conocimiento del nuevo proceso de parametrización se crea un manual donde se describe un paso a paso, para lograr tener una ejecución exitosa del proceso automatizado de parametrización, en el manual se expone la ruta origen donde se encuentra todo el desarrollo que se realizó, esto por si en futuros proyecto requieran su modificación.

El manual se agrega a el documento como el anexo 1 por lo extenso del mismo, este manual se socializo al área de Tecnología Planeación y Riesgo, con el fin de que expresaran sus dudas y ayudaran a mejorar el documentó, la versión que se anexa es la última después de las modificaciones que se sugirieron.

6.1 Resultados de la implementación

1) Se tuvo una mejora en los tiempos de parametrización en un 98.96%, ya que con el proceso anterior se tardaba una persona 4 o más días en una sola fuente, con el nuevo proceso solo se tarda 20 minutos, con estos dos datos se realiza una regla de tres.

   4 días = 32 horas laborales = 1920 Minutos

$$x = \frac{20m * 100}{1920m} = 1.04\%$$

   Esto repercute en una mejora en los Acuerdos de niveles de servicios, ya que con la anterior parametrización estos ANS se estaban incumpliendo en gran medida, ahora todas las fuentes de la semana quedan parametrizadas en 20 minutos no importa el número de campos.

2) Se redujeron los errores comunes a cero, gracias al nuevo proceso de parametrización ya que la mayoría de esto errores ocurría por el factor humano y con el nuevo proceso este factor es eliminado.
3) Se presentaba que los equipos que interfieren en las etapas superiores, realizaban cambios en la estructura del diccionario y en múltiples ocasiones decían que el error resultado de ese cambio era por parte de prototipado, gracias al nuevo proceso de parametria se puede evidenciar que no es así, ya que se cuenta con evidencia documentada que el error no fue en prototipado, ya que se copia la información tal como están en el diccionario origen.
4) Gracias a todos estos beneficios en varias ocasiones el jefe del área a expresado su conformidad con el nuevo proceso de parametrización, pues ya no se vencen los ANS de prototipado, además que los reprocesos por errores comunes se convirtieron en cero. Incitando a la innovación en los procesos que identifiquemos ineficiente.

Conclusiones

La automatización de la parametrización de fuentes, elimino los errores comunes que eran ocasionados por una persona al momento del parametrizado, lo que ayuda a mejor los tiempos de ANS y reducción el número de fuentes que son devueltas por presentar errores en el parametrizada. Además, se redujeron el número de tareas que se realizan en el prototipado de fuentes.

Con el proceso anterior una persona solo podía parametrizar una fuente a la vez y se demoraba dependiendo lo larga que fuera la fuente, gracias a la automatización de fuentes todas las fuentes de la semana se parametrizan de 5 a 20 minutos sin importar lo extensas que estas sean, una mejora que representa alrededor de 98.96% en productividad.

Lograr la automatización de fuentes marco un antes y un después, en el parametrizado que se venían dado en el banco, gracias a este desarrollo los tiempos de respuesta por parte de las personas de prototipado ahora son mejores ya que se quitó una de las actividades, dando una mejora con respecto a los otros departamentos que participan en la ingesta a el lago de latos.

Bibliografía

Bayrak, T. (2021). A framework for decision makers to design a business analytics platform for distributed organizations. *Technology in Society*, *67*, 101747. https://doi.org/10.1016/J.TECHSOC.2021.101747

Gökalp, M. O., Gökalp, E., Kayabay, K., Gökalp, S., Koçyiğit, A., & Eren, P. E. (2022). A process assessment model for big data analytics. *Computer Standards and Interfaces*, *80*. https://doi.org/10.1016/j.csi.2021.103585

Müller, O., Fay, M., & vom Brocke, J. (2018). The Effect of Big Data and Analytics on Firm Performance: An Econometric Analysis Considering Industry Characteristics. *Journal of Management Information Systems*, *35*(2), 488–509. https://doi.org/10.1080/07421222.2018.1451955

Olabode, O. E., Boso, N., Hultman, M., & Leonidou, C. N. (2022). Big data analytics capability and market performance: The roles of disruptive business models and competitive intensity. *Journal of Business Research*, *139*, 1218–1230. https://doi.org/10.1016/j.jbusres.2021.10.042

Pigni, F., Piccoli, G., & Watson, R. (2016). Digital Data Streams: Creating Value from the Real-Time Flow of Big Data. *California Management Review*, *58*(3), 5–25. https://doi.org/10.1525/cmr.2016.58.3.5

Samydurai, A., Revathi, K., Karthikeya, L., Vanathi, B., & Devi, K. (2022). An Enhanced Entity Model for Converting Relational to Non-Relational Documents in Hospital Management System Based on Cloud Computing. *IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India)*. https://doi.org/10.1080/02564602.2021.2016075

Shapiro, C. (1989). The Theory of Business Strategy. *The RAND Journal of Economics*, *20*(1), 125–137. https://doi.org/10.2307/2555656

Shmueli, G., & Koppius, O. R. (2011). Predictive Analytics in Information Systems Research. *MIS Quarterly*, *35*(3), 553–572. https://doi.org/10.2307/23042796

Yalcin, A. S., Kilic, H. S., & Delen, D. (2022a). The use of multi-criteria decision-making methods in business analytics: A comprehensive literature review. *Technological Forecasting and Social Change*, *174*, 121193. https://doi.org/10.1016/J.TECHFORE.2021.121193

Manual De Ejecución Del Parametrizado Automatizado Para Davivienda

Área Tecnología Planeación y Riesgo Del Banco Davivienda

Marzo de 2022

Contenido

1. Introducción

En este documento se realiza una descripción detallada del nuevo proceso de parametrización de fuentes. El cual se desarrolló en el interior del área de Tecnología Planeación y Riesgo del banco Davivienda, esto con el fin de mejorar los tiempos y los errores que se están cometiendo en la etapa de parametrizado de una fuente, repercutiendo en pérdida de tiempo y desgaste del personal con los reprocesos.

El manual se elabora con el fin de que cualquier integrante del área de Tecnología Planeación y Riesgo, pueda seguir el paso a paso y lograr llevar a cabo el proceso automatizado de parametrización de fuentes nuevas y con ajustes en el diccionario, para cada una de las casuísticas van a tener un numero de pasos que debe seguir, muchos de estos pasos se repiten.

2. Objetivo

Generar un manual para el área de Tecnología Planeación y Riesgo del banco Davivienda que describa el paso a paso de la utilización del nuevo proceso de parametrización de una fuente.

3. Fase De Pre-Alistamiento Para Fuentes Nuevas

Para parametrizar una fuente nueva se van a dirigir al siguiente link: https://docs.google.com/spreadsheets/d/1vFKWsL616yGmqzp08ALJaSAPmXXKAe9cmyEd-UtWAQc/edit#gid=1706855886.

En el link se va visualizar el ***Tablero Control Gobierno de Datos***, se van a dirigir a la pestaña ***Planeado Tablas TI***, una vez en la pestaña realizaran un ***filtro en la columna X*** (Estado TI), seleccionan las etapas: ***En Proceso de prototipado, En Proceso de Integración y Cierre de documentación***, Estas son etapas que pertenecen al área. En la imagen 1 se puede ver un ejemplo.

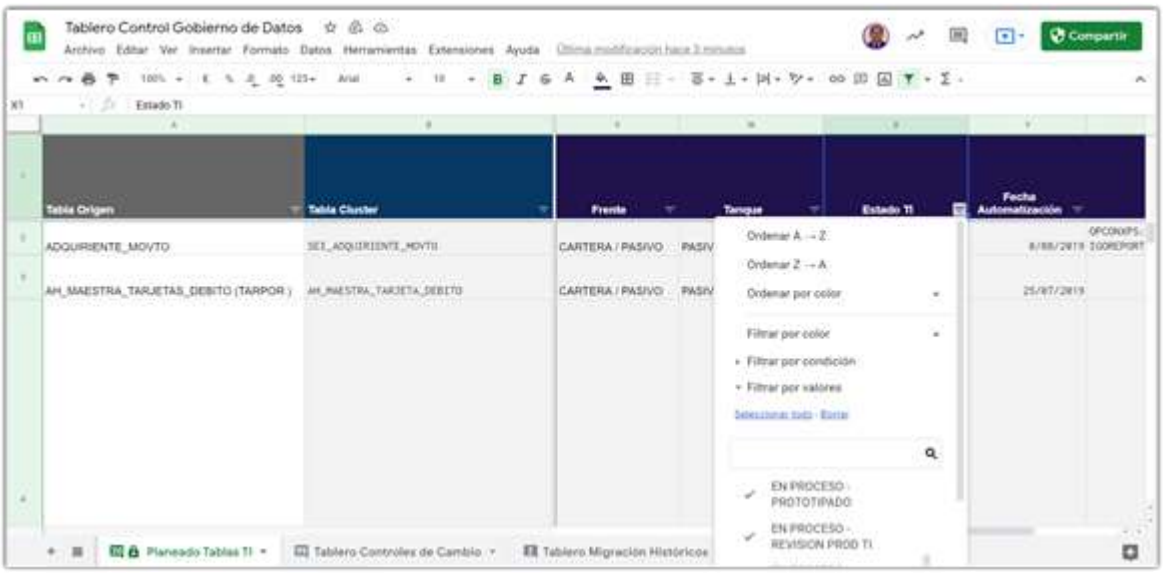


Imagen 1: Tablero de control

Proceden a ***copiar las columnas A (Tabla Origen) y B (Tabla Clúster)***, y pegan la información de esas dos columnas en otro Excel el cual debe llevar por nombre ***PARAMETROS_TI_INGESTA_AAAAMMDD***, como se muestra en la imagen 2.

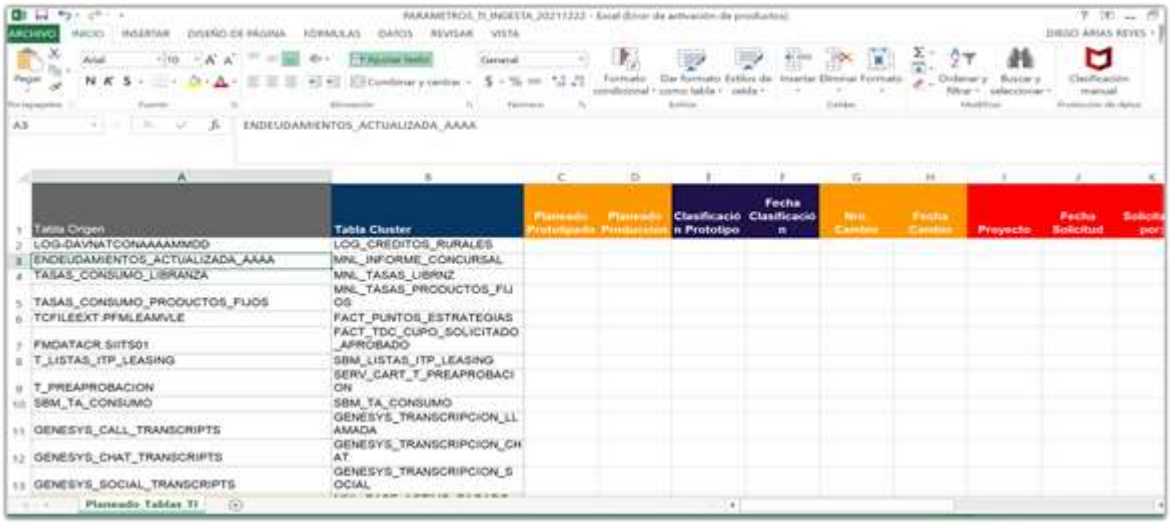


Imagen 2: Tabla Parámetros TI

Guardan el nuevo Excel con la información y lo suben a la carpeta ***odifiles*** del servidor de intercambio, luego ejecutan su **script** (Shell) en este ejemplo: ***Ejecucion_Limpieza_Jesus.sh***

para la el Excel: ***PARAMETROS_TI_INGESTA_AAAAMMDD***, una vez ejecutado el script se realiza el cargue a Cloudera, por medio de HUE.

Una vez cargada la información terminan la fase de pre-alistamiento, esta fase es fundamental para los siguientes pasos ya que la ejecución del Workflows va a buscar esta información que está en el documento. La ejecución de los Querys se realizará en el editor **Impala** de HUE.

Paso 1: Buscar en Cloudera la carpeta que contiene todo el desarrollo.

Ruta:

- /
- workflows/
- prototipado/
- 8096_TEC_PLANEACION_RIESGO/
- **piloto_automatizacion_prototipado_aprovisionamiento**

Imagen 3: Carpeta del piloto

Nota: En esta carpeta van a encontrar todo el desarrollo que se ha realizado, si se requiere modificar el código abrirán él **.py** y editan lo que ustedes quieran, si van hacer esto realizar un backup de ese archivo que esta 100% funcional.

Paso 2: Seleccionan el recuadro de workflow.xml

Imagen 3: Carpeta del piloto

Paso 3: Envían el workflow.xml

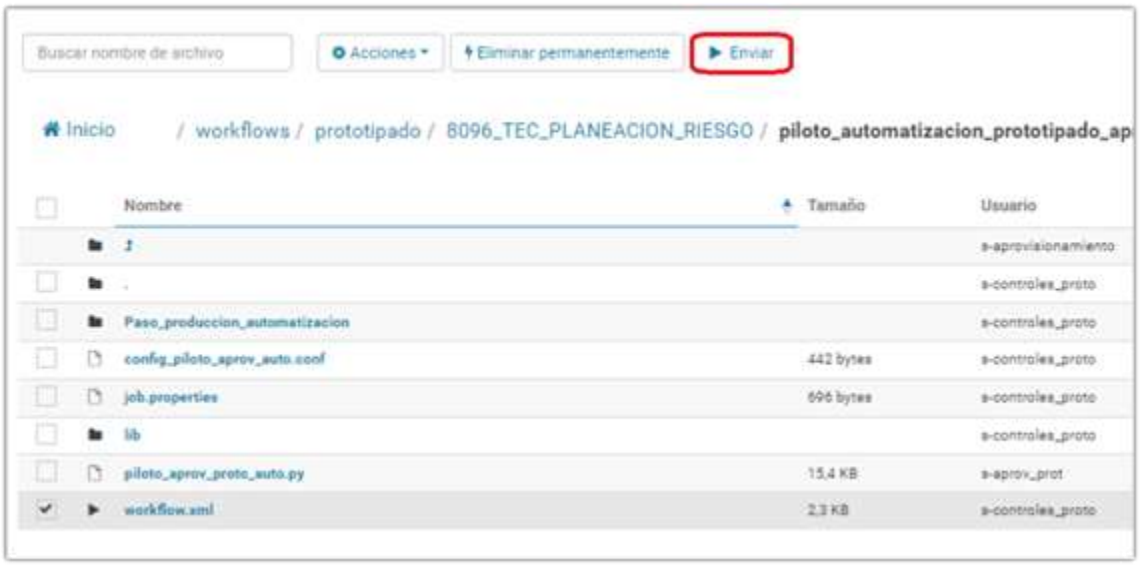


Imagen 4: Enviar workflow

Se generará una pestaña emergente dan clic nuevamente en **Enviar,** no editan nada**.**

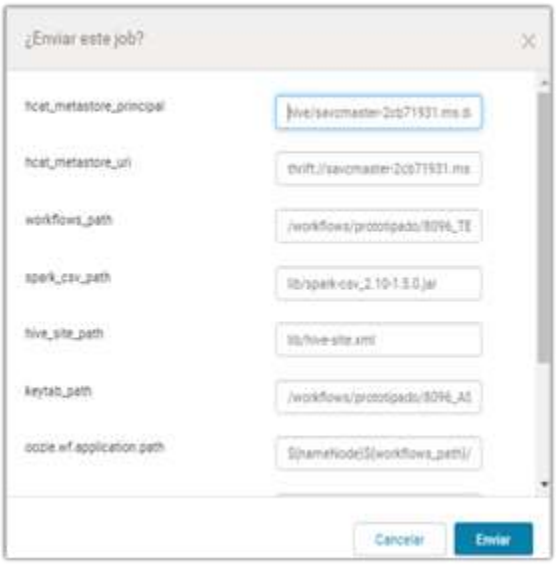


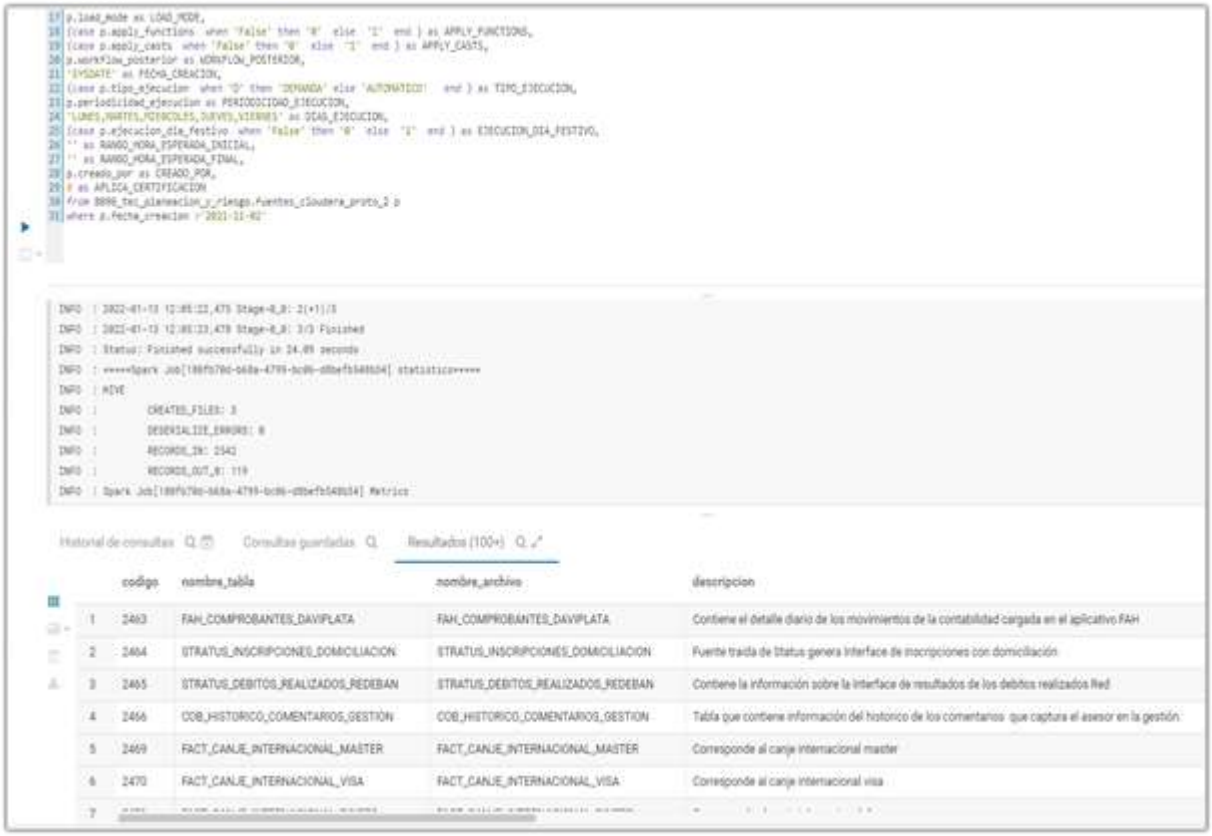
Imagen 5: Ventana emergente

3.1 Para Fuentes Nuevas En Fuentes_Cloudera

Paso 4: Realizan la siguiente consulta:

```
select
p.codigo as CODIGO,
p.nombre_tabla as NOMBRE_TABLA,
p.nombre_archivo as NOMBRE_ARCHIVO,
p.descripcion as DESCRIPCION,
p.delimitador_campo as DELIMITADOR_CAMPO,
(case p.campos_particion  when 'PERIODO' then '1'  else  '0'  end ) as
TABLA_PARTICIONADA,
p.periodicidad as PERIODICIDAD,
(case p.activo  when 'False' then '0'  else  '1'  end ) as ACTIVO,
'APP' as PLATAFORMA,
'APP' as APLICATIVO,
'DAT9' as DATABASDE,
p.sublocation as SUBLOCATION,
p.formato as FORMATO,
(case p.header  when 'False' then '0'  else  '1'  end ) as HEADER,
p.load_mode as LOAD_MODE,
(case p.apply_functions  when 'False' then '0'  else  '1'  end ) as APPLY_FUNCTIONS,
(case p.apply_casts  when 'False' then '0'  else  '1'  end ) as APPLY_CASTS,
p.workflow_posterior as WORKFLOW_POSTERIOR,
'SYSDATE' as FECHA_CREACION,
(case p.tipo_ejecucion  when 'D' then 'DEMANDA' else 'AUTOMATICO'   end ) as
TIPO_EJECUCION,
p.periodicidad_ejecucion as PERIODICIDAD_EJECUCION,
'LUNES,MARTES,MIERCOLES,JUEVES,VIERNES' as DIAS_EJECUCION,
(case p.ejecucion_dia_festivo  when 'False' then '0'  else  '1'  end ) as
EJECUCION_DIA_FESTIVO,
'' as RANGO_HORA_ESPERADA_INICIAL,
'' as RANGO_HORA_ESPERADA_FINAL,
p.creado_por as CREADO_POR,
0 as APLICA_CERTIFICACION
from 8096_tec_planeacion_y_riesgo.fuentes_cloudera_proto_2 p
where p.fecha_creacion >'2021-11-02'
```

Lo único que van a modificar de esta consulta es la fecha de creación, esta resaltada en amarrillo. El resultado del paso 4 es la imagen 6.



	codigo	nombre_tabla	nombre_archivo	descripcion
1	2463	FAH_COMPROBANTES_DAVIPLATA	FAH_COMPROBANTES_DAVIPLATA	Contiene el detalle diario de los movimientos de la contabilidad cargada en el aplicativo FAH
2	2464	STRATUS_INSCRIPCIONES_DOMICILIACION	STRATUS_INSCRIPCIONES_DOMICILIACION	Fuente traida de Status genera Interface de inscripciones con domiciliación
3	2465	STRATUS_DEBITOS_REALIZADOS_REDEBAN	STRATUS_DEBITOS_REALIZADOS_REDEBAN	Contiene la información sobre la interface de resultados de los debitos realizados Red
4	2466	COB_HISTORICO_COMENTARIOS_GESTION	COB_HISTORICO_COMENTARIOS_GESTION	Tabla que contiene información del historico de los comentarios que captura el asesor en la gestión
5	2469	FACT_CANJE_INTERNACIONAL_MASTER	FACT_CANJE_INTERNACIONAL_MASTER	Corresponde al canje internacional master
6	2470	FACT_CANJE_INTERNACIONAL_VISA	FACT_CANJE_INTERNACIONAL_VISA	Corresponde al canje internacional visa

Imagen 6: Resultado consulta para fuentes Cloudera

Paso 5: Descargan el resultado de la consulta en un archivo CSV, como se muestra a continuación:

Imagen 7: Descargar CSV

Paso 6: El archivo .CSV lo abren en Notepad++, Copian la información **sin los títulos** y se dirigen a la plantilla de Excel, la cual ya tiene los títulos que no copiaron, pegan la información en la hoja **fuentes_cloudera.** link plantilla:

https://docs.google.com/spreadsheets/d/13mV0gymFOz3GodFZrRRlg73EOUTMmpw8/edit#gid=664570380 Les debe quedar como en la imagen 8.

Imagen 8: Plantilla Fuentes_cloudera

Paso 7: Deben seleccionar los datos, los cuales pegaron en el anterior paso, una vez seleccionados se dirigen a la pestaña ***Datos*** del Excel y dan clic en ***dividir texto en columnas***. En la imagen 9 van observar una explicación grafica de lo que deben realizar.

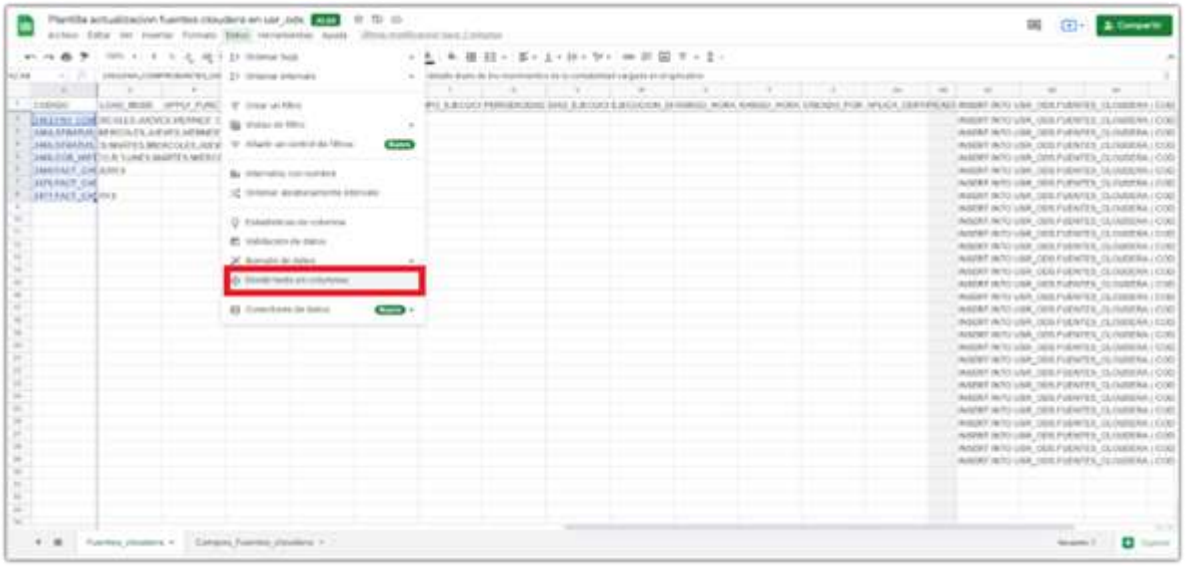


Imagen 9: Dividir texto

Si completaron el paso 7 de manera correcta les debe quedar como en la imagen 10, en el la cual ya pueden observar el despliegue de toda la información que estaba en la columna A.

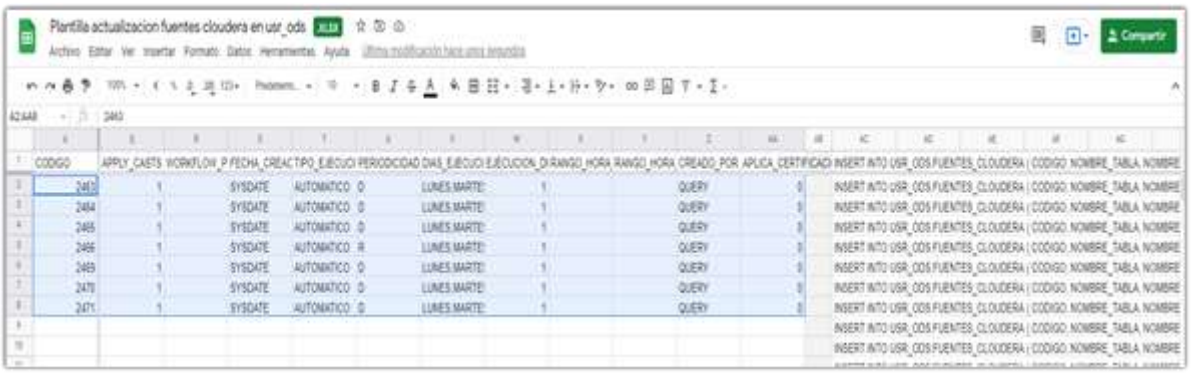


Imagen 10: Resultado de la división del texto

Paso 8: Proceden a copiar la desde la columna **AC** del Excel, donde ya esta preparado el insert con la información que acaban de ordenar en el Excel, copian y se dirigen al SQL Developer, pegan todos lo insert. El resultado del paso 8 debe ser como la imagen 11.

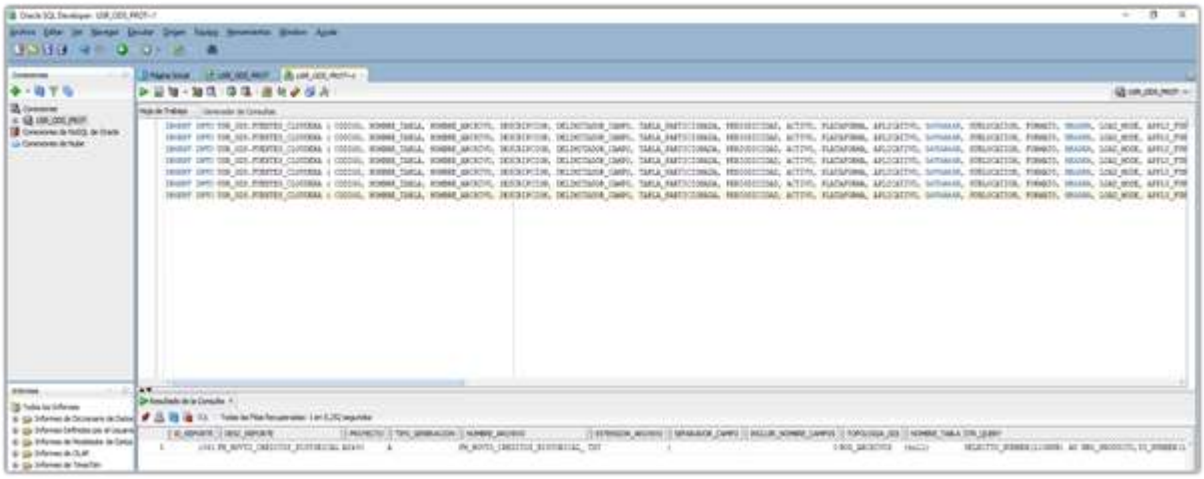


Imagen 11: SQL Developer

Paso 9: Seleccionan todos los registros que pegaron el SQL y dan clic en **ejecutar**  , Se seleccionan para ejecutar todos los registros. Ejemplo de lo que deben realizar en el paso 9 imagen 12.

Imagen 12: Ejecutan las sentencias

Paso 10: Darán clic en el botón de **confirmar,** para que se guarden los cambios realizados a la tabla fuentes Cloudera.

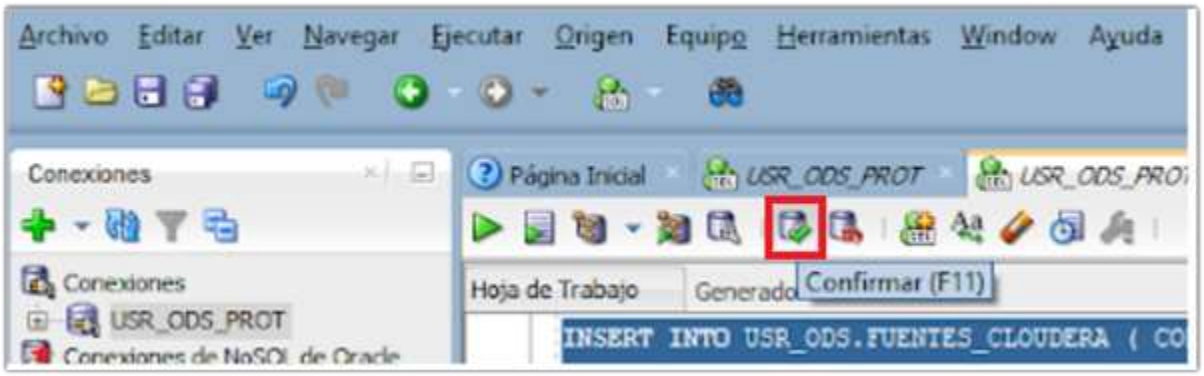


Imagen 13: Confirman

3.2 Para Fuentes Nuevas En Campos_Fuentes

Una vez terminan los pasos para fuentes Cloudera deben continuar con los pasos que estarán modificando la información de la tabla de campos fuentes Cloudera, Esta tabla contiene la información de todos los campos de las fuentes que serán creadas en el lago de datos.

Se continua la secuencian de los pasos porque es un solo proceso.

Paso 11: Realizan la siguiente consulta:

```
select
c.codigo_fuente as CODIGO_FUENTE ,
c.consecutivo_campo as CONSECUTIVO_CAMPO ,
upper(c.nombre_campo) as NOMBRE_CAMPO ,
c.descripcion as DESCRIPCION ,
(case c.campo_particion  when 'False' then '0'  else  '1'  end ) as
CAMPO_PARTICION ,
(case c.campo_llave_primaria  when 'False' then '0'  else  '1'  end ) as
CAMPO_LLAVE_PRIMARIA ,
c.tipo_campo as TIPO_CAMPO ,
 replace(c.longitud_campo,'.',',') as longitud_campo ,
nvl(c.opcional,0) as OPCIONAL ,
c.fx_campo as FX_CAMPO ,
c.comentarios as COMENTARIOS ,
0  as ORDEN_CAMPO_PARTICION
from 8096_tec_planeacion_y_riesgo.campos_fuentes_cloudera_PROTO c
LEFT JOIN 8096_tec_planeacion_y_riesgo.fuentes_cloudera_proto_2 p
ON  CAST( p.codigo AS INT )= CAST(c.codigo_fuente AS INT )
where p.fecha_creacion >'2021-11-02'
union all
select
c.codigo_fuente as CODIGO_FUENTE ,
MAX(c.consecutivo_campo)+1 as CONSECUTIVO_CAMPO ,
p.campos_particion as NOMBRE_CAMPO ,
p.campos_particion as DESCRIPCION ,
'1' as CAMPO_PARTICION ,
'0' as CAMPO_LLAVE_PRIMARIA ,
'NUMBER' as TIPO_CAMPO ,
'8' as LONGITUD_CAMPO ,
0 as OPCIONAL ,
'' as FX_CAMPO ,
'' as COMENTARIOS ,
0 as ORDEN_CAMPO_PARTICION
from 8096_tec_planeacion_y_riesgo.campos_fuentes_cloudera_PROTO c
   , 8096_tec_planeacion_y_riesgo.fuentes_cloudera_proto_2 p
where p.codigo = c.codigo_fuente
and  p.fecha_creacion >'2021-11-02'
```

```
and p.campos_particion <>"
GROUP BY c.codigo_fuente  ,
p.campos_particion
```

Antes de ejecutar el código deben modificar la fecha de creación, esta sombreada de color amarillo para facilitar su reconocimiento. Una vez ejecutado el select, el resultado es el que se muestra en la imagen 14.

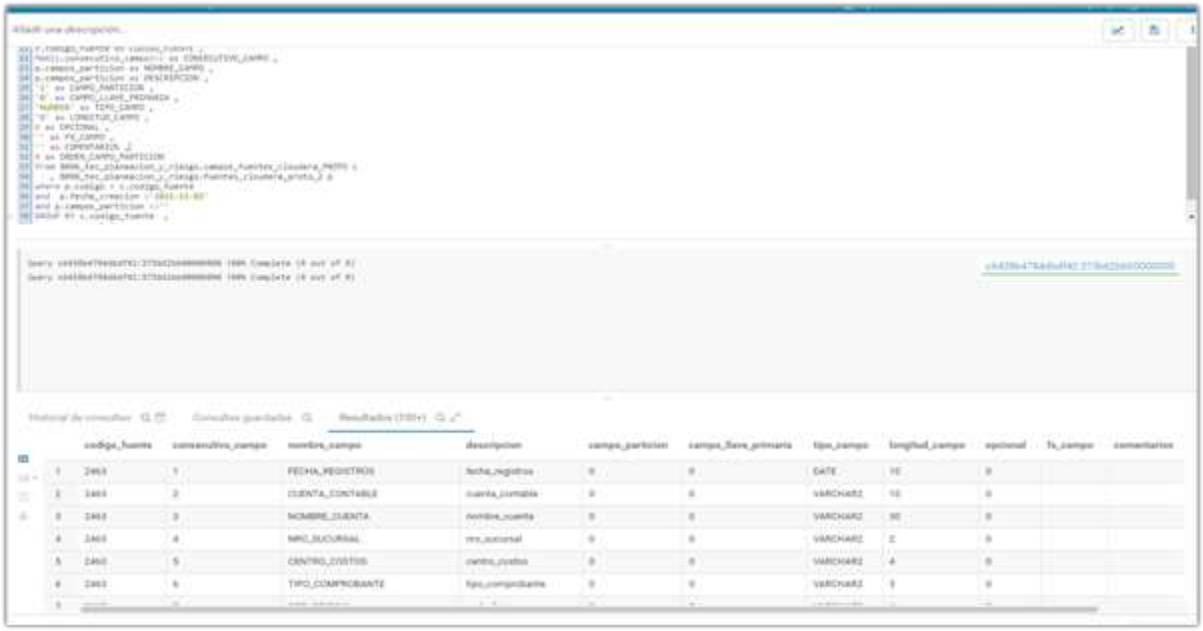


Imagen 14: Resultado de la consulta campos fuentes

Paso 12: Descargan el resultado de la consulta en un archivo csv, como se muestra en la imagen 15.

Imagen 15: Descarga CSV campos fuentes

Paso 13: Abren el archivo CSV en un Notepad++, copian los registros **sin los títulos** y se dirigen a la plantilla de Excel y pegan en la hoja **campos_fuente_cloudera.** Link plantilla: https://docs.google.com/spreadsheets/d/13mV0gymFOz3GodFZrRRlg73EOUTMmpw8/edit#gid=664570380, Resultado del paso 13 imagen 16.

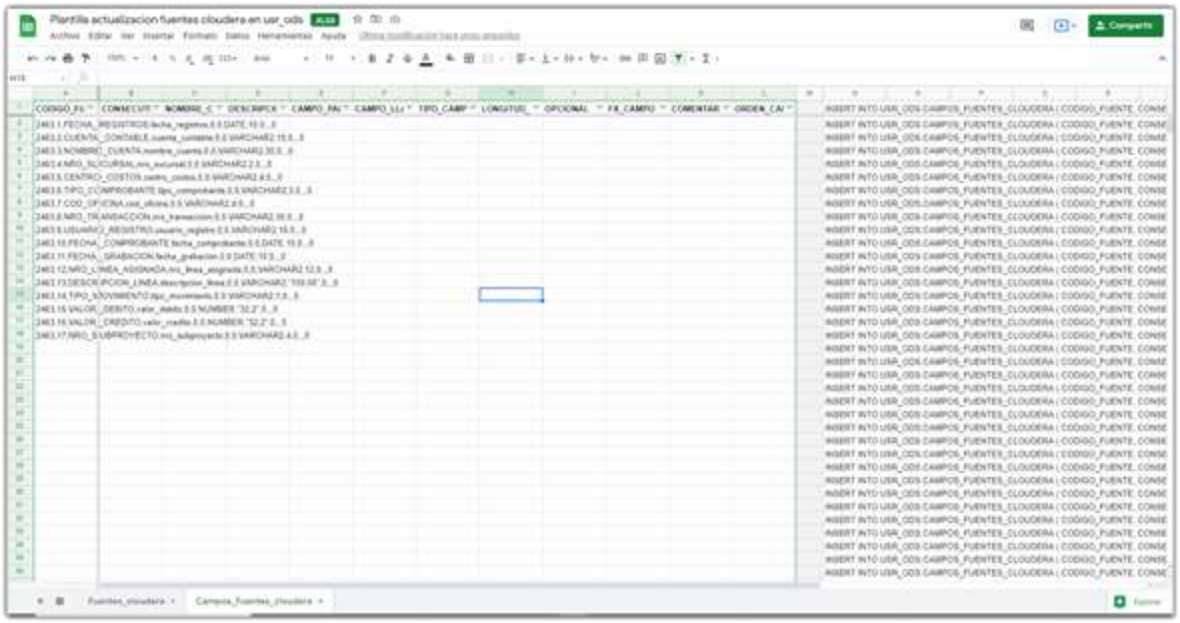


Imagen 16: Plantilla para campos fuentes

Paso 14: Deben seleccionar los datos, los cuales pegaron en el anterior paso, una vez seleccionados se dirigen a la pestaña ***Datos*** del Excel y dan clic en ***dividir texto en columnas***. En la imagen 17 van observar una explicación grafica de lo que deben realizar.

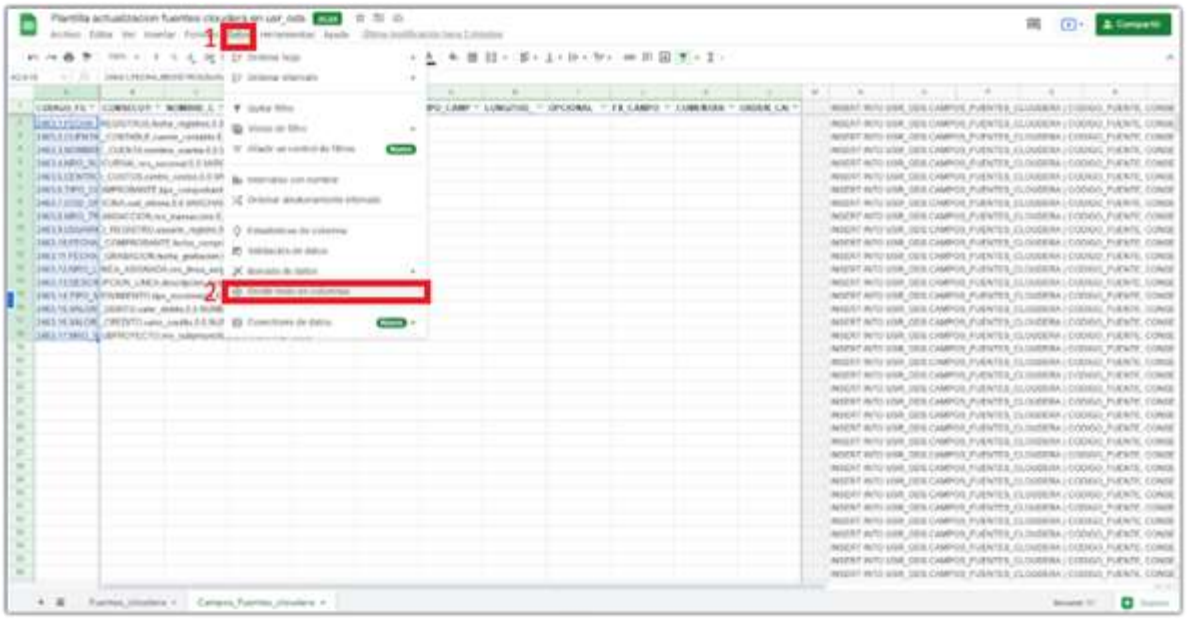


Imagen 17: División de texto campos fuentes

Si completaron el paso 14 de manera correcta les debe quedar como en la imagen 18, en el la cual ya pueden observar el despliegue de toda la información que estaba en la columna A.

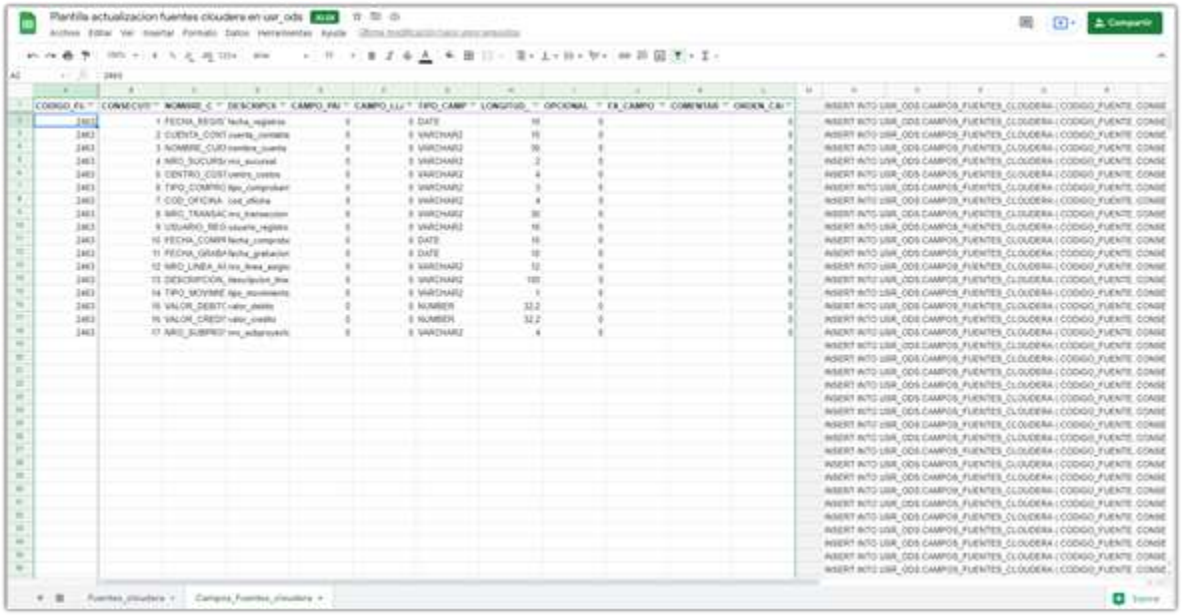


Imagen 18: Resultado de la división campos fuentes

Paso 15: Proceden a copiar la información desde la columna **N** hasta donde tengamos datos en las columnas, Estos son los insert ya con la información que acaban de ordenar en el Excel, copian y se dirigen al SQL Developer, pegan todos lo insert. El resultado del paso 15 debe ser como la imagen 19.

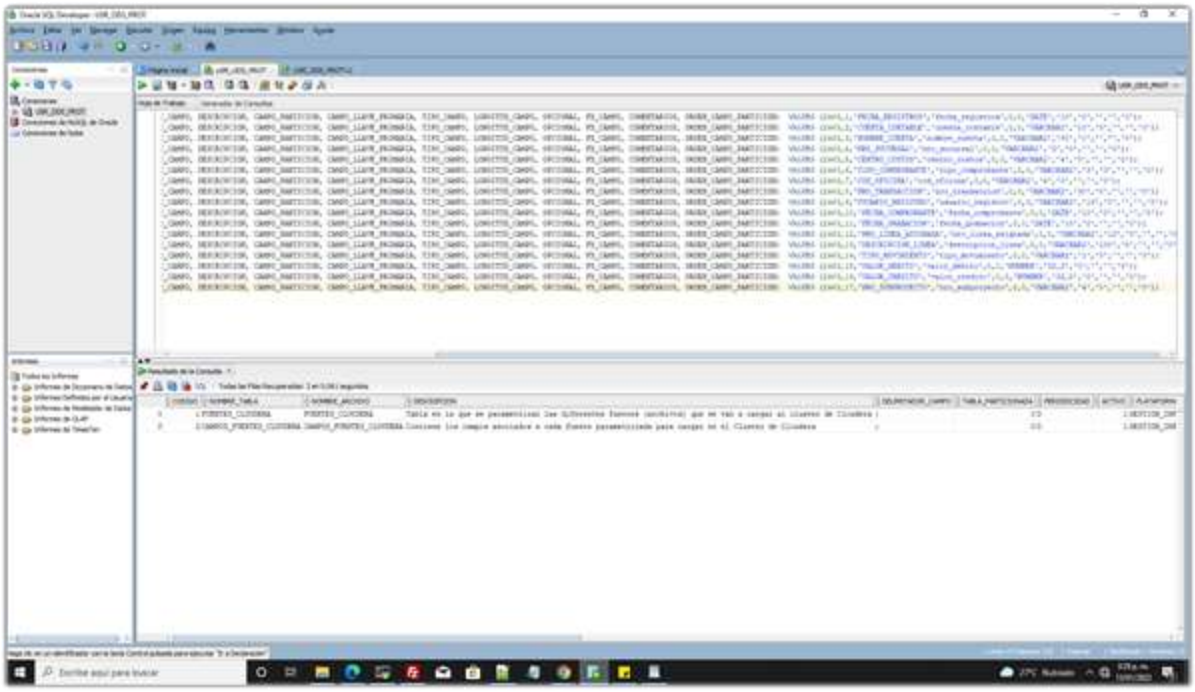


Imagen 19: SQL Developer Campos fuentes

Paso 16: Seleccionan todos los registros que pegaron el SQL y dan clic en **ejecutar**  , Se seleccionan para ejecutar todos los registros. Ejemplo de lo que deben realizar en el paso 16 imagen 20.

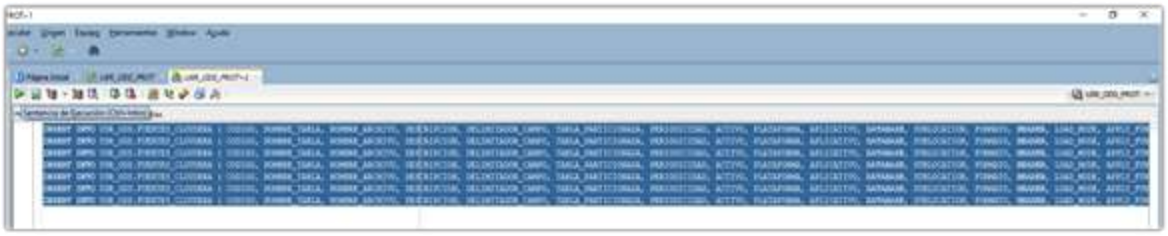


Imagen 20: Ejecución de los insert campos fuentes

Paso 17: Como paso de creación de una fuente nueva darán clic en el botón de **confirmar,** Para así confirmar los cambios en la tabla en la base de datos.

Imagen 21: Confirmación de los insert

El paso 17 es el ultimo que realizan para la creación de una fuente nueva, una vez terminado este paso puede realizar un select, consultando las fuentes nuevas que acabaron de crear y comprobar sus campos.

4. Fase De Pre-Alistamiento Para Fuentes Con Ajustes En El Diccionario

Fuentes Con Novedades:
Abrirán el archivo ***TABLANOVEDAD_AAAAMMDD*** este es un Excel donde van a llenar los campos que son requeridos para las fuentes que contiene novedades: ***Tabla Origen, Tabla Clúster, NOMBRE_CAMPO y CAMPO_NOVEDADES***. Imagen 22.

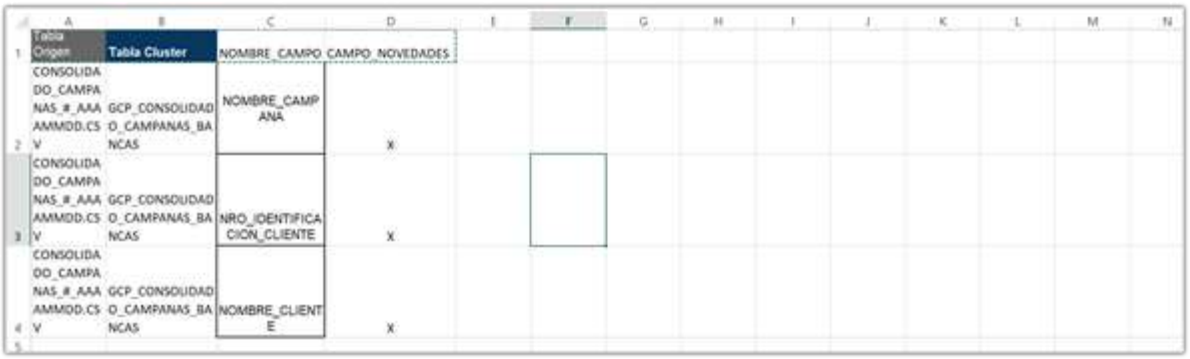


Imagen 22: Excel novedades

Guardan el archivo en el servidor de intercambio ***odifiles***, Ejecutan su **script** en este caso es: **Ejecucion_Limpieza_Jesus.sh**, para el Excel ***TABLANOVEDAD_AAAAMMDD***, una vez ejecutado el script se realiza el cargue a Cloudera, por medio de HUE.

Fuentes Con Cambios:
Para realizar cambios en una fuente que ya está creada se dirigen a link: https://docs.google.com/spreadsheets/d/1vFKWsL616yGmqzp08ALJaSAPmXXKAe9cmyEd-UtWAQc/edit#gid=1706855886, es el mismo Excel de la imagen 1, este caso se van a dirigir a la pestaña ***Tablero Controles de Cambio*** y van a realizar un **filtro en la columna I** (Estado) por la etapa ***En Proceso de Prototipado***.

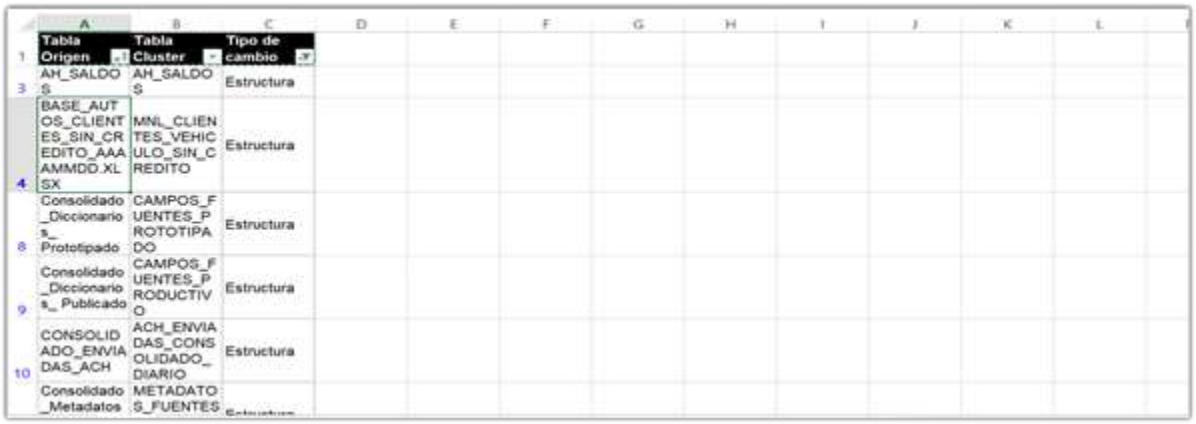


Imagen 23: Tabla cambios

Proceden a copiar la información de las columnas A, B y C y lo pegan en un Excel como se muestra en la imagen 23 el cual debe llevar por nombre ***TABLA_NOVEDADES_AAAAMMDD,*** Guardan el archivo en el servidor de intercambio en la carpeta ***odifiles***, luego ejecutan su ***script*** en este caso es: ***Ejecucion_Limpieza_Jesus.sh,*** para el archivo: *TABLA_NOVEDADES_AAAAMMDD.*

4.1 Para Fuentes Con Ajustes En El Diccionario Fuentes_Cloudera

Paso 1: Van a realizar los pasos 1, 2 y 3 realizados para fuentes nuevas, estos pasos son iguales para fuentes con cambio o novedades.

Paso 2: Ejecutan el siguiente query, el cual va dar como resultado el código de las fuentes de controles de cambios o novedades, esto gracias a el paso de pre-alistamiento.

```
select
distinct
p.codigo as CODIGO
from 8096_tec_planeacion_y_riesgo.fuentes_cloudera_proto_2 p
where p.creado_por like '%QUERY 2021%'
```

Del query solo van a modificar la fecha de creación. Una vez lo ejecuten el resultado será algo similar a la imagen 24, en este ejemplo solo una fuente será modificada, por lo tanto solo se tiene un código de fuente.

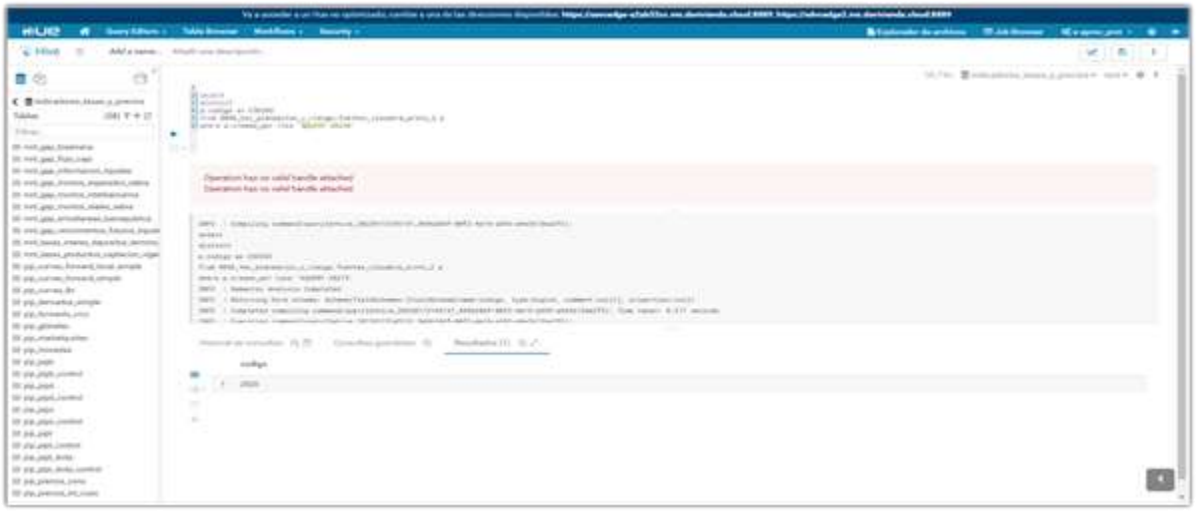


Imagen 24: Resultado novedades o cambios

Paso 3: Se dirigen a SQL Developer y pegan la siguiente línea de código.
DELETE FROM USR_ODS.FUENTES_CLOUDERA WHERE CODIGO IN (2520).
Remplazan los números dentro de los paréntesis por los del resultado del paso 2, ejemplo si se tienen varias fuentes (2520,3032……). Este paso esta borrando las fuentes de la tabla de ***fuentes_cloudera***.

Paso 4: Proceden a ejecutar el siguiente query:

```
select
p.codigo as CODIGO,
p.nombre_tabla as NOMBRE_TABLA,
p.nombre_archivo as NOMBRE_ARCHIVO,
p.descripcion as DESCRIPCION,
p.delimitador_campo as DELIMITADOR_CAMPO,
(case p.campos_particion  when 'PERIODO' then '1'  else  '0'  end ) as
TABLA_PARTICIONADA,
```

```
p.periodicidad as PERIODICIDAD,
(case p.activo  when 'False' then '0'  else  '1'  end ) as ACTIVO,
'APP' as PLATAFORMA,
'APP' as APLICATIVO,
'DAT9' as DATABASDE,
p.sublocation as SUBLOCATION,
p.formato as FORMATO,
(case p.header  when 'False' then '0'  else  '1'  end ) as HEADER,
p.load_mode as LOAD_MODE,
(case p.apply_functions  when 'False' then '0'  else  '1'  end ) as
APPLY_FUNCTIONS,
(case p.apply_casts  when 'False' then '0'  else  '1'  end ) as APPLY_CASTS,
p.workflow_posterior as WORKFLOW_POSTERIOR,
'SYSDATE' as FECHA_CREACION,
(case p.tipo_ejecucion  when 'D' then 'DEMANDA' else 'AUTOMATICO'   end ) as
TIPO_EJECUCION,
p.periodicidad_ejecucion as PERIODICIDAD_EJECUCION,
'LUNES,MARTES,MIERCOLES,JUEVES,VIERNES' as DIAS_EJECUCION,
(case p.ejecucion_dia_festivo  when 'False' then '0'  else  '1'  end ) as
EJECUCION_DIA_FESTIVO,
'' as RANGO_HORA_ESPERADA_INICIAL,
'' as RANGO_HORA_ESPERADA_FINAL,
p.creado_por as CREADO_POR,
0 as APLICA_CERTIFICACION
from 8096_tec_planeacion_y_riesgo.fuentes_cloudera_proto_2 p
where p.creado_por like '%QUERY 2021%'
```

Una vez más solo modifican la **fecha** que esta sombreada de color amarillo. Como resultado del paso 4 van a tener al similar a la imagen 25.

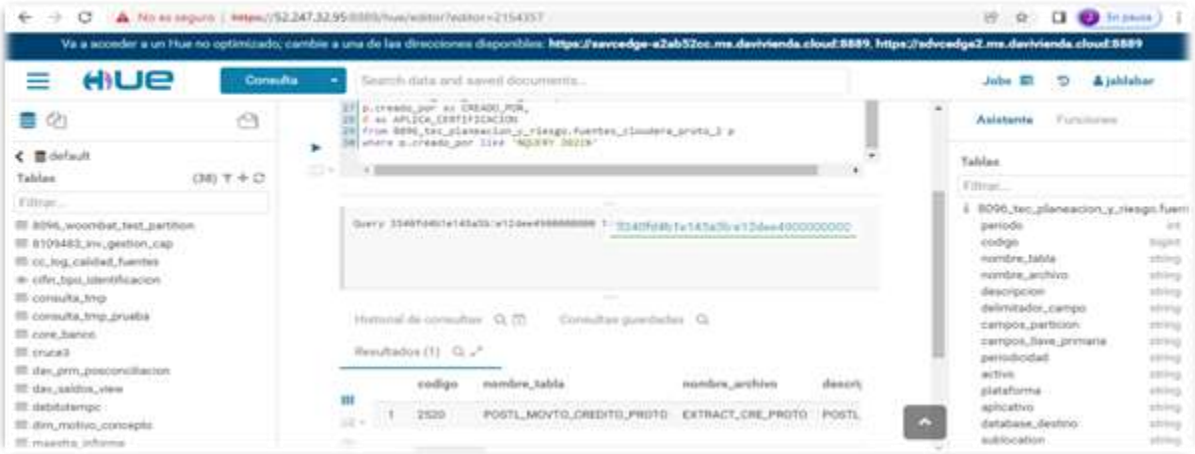


Imagen 25: Resultado query novedades o cambios

Paso 5: Los siguientes es repetir desde el **paso 5 al 10** del procedimiento **Para fuentes nuevas en fuentes_cloudera.**

4.2 Para Fuentes Con Ajustes En El Diccionario Campos_Fuentes_Cloudera

Paso 6: se toma la información obtenida en el **paso 2 de fuentes con ajustes** en el diccionario o novedades.

Paso 7: Se dirigen a SQL Developer y pegan la siguiente línea de código.
DELETE FROM USR_ODS.CAMPOS_FUENTES_CLOUDERA WHERE CODIGO_FUENTE IN (2520). Como en el paso 3 de fuentes con ajustes solo van editar los códigos del paréntesis. **En este paso están realizando el borrador de la fuente que se va ajustar, pero es para la tabla *campos fuentes Cloudera***.

Paso 8: Proceden a ejecutar la siguiente consulta:

```
select
c.codigo_fuente as CODIGO_FUENTE ,
c.consecutivo_campo as CONSECUTIVO_CAMPO ,
upper(c.nombre_campo) as NOMBRE_CAMPO ,
c.descripcion as DESCRIPCION ,
(case c.campo_particion  when 'False' then '0'  else  '1'  end ) as
CAMPO_PARTICION ,
(case c.campo_llave_primaria  when 'False' then '0'  else  '1'  end ) as
CAMPO_LLAVE_PRIMARIA ,
c.tipo_campo as TIPO_CAMPO ,
 replace(c.longitud_campo,'.',',') as longitud_campo ,
nvl(c.opcional,0) as OPCIONAL ,
c.fx_campo as FX_CAMPO ,
c.comentarios as COMENTARIOS ,
0  as ORDEN_CAMPO_PARTICION
from 8096_tec_planeacion_y_riesgo.campos_fuentes_cloudera_PROTO c
LEFT JOIN 8096_tec_planeacion_y_riesgo.fuentes_cloudera_proto_2 p
ON  CAST( p.codigo AS INT )= CAST(c.codigo_fuente AS INT )
where p.creado_por like '%QUERY 2021%'
union all
select
c.codigo_fuente as CODIGO_FUENTE ,
MAX(c.consecutivo_campo)+1 as CONSECUTIVO_CAMPO ,
p.campos_particion as NOMBRE_CAMPO ,
p.campos_particion as DESCRIPCION ,
'1' as CAMPO_PARTICION ,
'0' as CAMPO_LLAVE_PRIMARIA ,
'NUMBER' as TIPO_CAMPO ,
```

```
'8' as LONGITUD_CAMPO ,
0 as OPCIONAL ,
'' as FX_CAMPO ,
'' as COMENTARIOS ,
0 as ORDEN_CAMPO_PARTICION
from 8096_tec_planeacion_y_riesgo.campos_fuentes_cloudera_PROTO c
  , 8096_tec_planeacion_y_riesgo.fuentes_cloudera_proto_2 p
where p.codigo = c.codigo_fuente
and  p.creado_por like '%QUERY 2021%'
and p.campos_particion <>''
GROUP BY c.codigo_fuente  ,
p.campos_particion
```

Una vez más solo modifican la **fecha**.

Paso 9: Como últimos pasos va a repetir los pasos desde el **5 al 10** del procedimiento **Para fuentes nuevas en *campos_fuentes_cloudera.***