Universidad de Guantánamo

Universidad de Guantánamo

Facultad de Ingeniería y Ciencias Técnicas

Ingeniería Informática

Tesis presentada en opción al Título de Ingeniero Informático

Perfilado y mapeo de datos de las fuentes de datos del delito en el MININT Guantánamo.

Autor: Adrián Brown Lewis

Tutores: Ms. C. Carla María Alonso Jane

Ing. 1er Tte. Eudel Campuzano Fuentes

Guantánamo, junio 2020

DECLARACIÓN DE AUTORÍA

El que suscribe, Adrián Brown Lewis, hago constar que el trabajo titulado Perfilado y mapeo de datos de las fuentes de datos del delito en el MININT Guantánamo, fue realizado como parte de la culminación de los estudios de la especialidad de Ingeniería Informática, autorizando a que el mismo sea utilizado por la institución, para los fines que estime conveniente, tanto de forma parcial como total y que además no podrá ser presentado en eventos ni publicado sin la autorización de la Universidad.

Para que así conste firmo la presente a los _____ días del mes de _____ del año 2020.

_____	_____
Adrián Brown Lewis	MsC. Carla María Alonso Jane
Firma del Autor	Firma del Tutor

Dedicatoria

Dedico este trabajo al sistema educativo cubano por permitirme el tránsito por todos sus niveles y llegar a este momento cumbre de mis estudios. Asimismo a todos los profesores que participaron de forma activa en mi formación profesional y muy en especial a la Revolución por haber concebido el proyecto social que permitió la concepción organizativa e inclusiva para alcanzar la formación profesional de sus miembros.

Agradecimientos

Primeramente, me gustaría agradecerle a mi madre, quien fue mi guía y apoyo en todo momento de mi vida, y en este proceso tan hermoso, pero a la vez tan difícil, gracias a mí que todo lo puede. Gracias a mi familia amada, por todo lo que me ha brindado durante mis 24 años de vida, principalmente por su incondicional amor. A mis tutores que acompañaron todos mis pasos, no solo como profesores, sino también como amigos incondicionales y además a todas las personas que de una forma u otra me aconsejaron para llegar a feliz término.

Resumen

El órgano de informática y comunicaciones en el MININT de Guantánamo dispone de varias aplicaciones que recopilan datos relacionados en los delitos. Cada una de ella dispone de su propia base de datos y no interactúan entre sí. Este órgano desea realizar un datamart para analizar los patrones del delito por lo que debe realizar un proceso de extracción, transformación y carga de datos de las fuentes de datos primarias a un almacenamiento intermedio que los unifique y luego al datamart. Por la complejidad de este proceso se realiza un perfilado y mapeo de los datos desde las fuentes de datos al almacenamiento intermedio. Además se diseña de manera eficiente como debe ser el almacenamiento intermedio. Con el perfil y mapeo de datos y el diseño del almacenamiento intermedio se podrá realizar de manera correcta y eficiente el proceso de extracción, transformación y carga de los datos.

Palabras claves: datamart; perfil de datos; mapeo de datos delitos; fuentes de datos, extracción, transformación y carga de los datos; patrones del delito

Abstract

The computer and communications body at the MININT in Guantánamo has several applications that collect data related to crimes. Each of them has its own database and does not interact with each other. This body wants to carry out a datamart to analyze the patterns of crime, so it must carry out a process of extraction, transformation and loading of data from the primary data sources to an intermediate storage that unifies them and then to the datamart. Due to the complexity of this process, a profiling and mapping of the data is performed from the data sources to the intermediate storage. It is also efficiently designed as the buffer should be. With the profile and mapping of data and the design of the intermediate storage, the process of data extraction, transformation and loading can be carried out correctly and efficiently.

Keywords: datamart; data profile; crime data mapping; data sources, data extraction, transformation and loading; crime patterns

ÍNDICE

INTRODUCCIÓN

En la sociedad actual ser víctima de un delito es uno de los mayores peligros a los que está sometida una persona. Las estadísticas demuestran que en cada región del mundo el delito incrementa por año, dejando tras sí una ola de daños que golpea fuertemente al sistema social de cada país. El continente americano es uno de los más violentos del mundo. Ocupa en el año 2019 el segundo escaño por homicidios con un 15,4 % por cada 100 000 habitantes, superando la media de homicidios a nivel global (6,9%).

La identificación y análisis táctico de patrones delictuales es una responsabilidad primordial de los analistas de las agencias policiales alrededor del mundo. Cada día, los analistas consultan y buscan datos en esfuerzo por vincular casos a través de factores claves y diseminar información acerca de patrones reconocidos y recién descubiertos a otras agencias policiales. Este análisis mejora la seguridad de las comunidades, facilitando la respuesta de la policía, que puede a su vez, prevenir y reducir la delincuencia.

El análisis de patrón de delito es una técnica de minería de datos la cual se centra en los componentes de los delitos para encontrar similitudes entre ellos. Esta técnica es empleada para extraer e identificar información útil que se convierte en conocimiento a partir de grandes bases de datos, data warehouses o datamart.

En Cuba el Ministerio del Interior (MININT), es el encargado de prevenir, neutralizar y esclarecer las actividades delictivas de carácter común y preservar el orden público y la seguridad ciudadana. Siendo la encargada directa de los casos delictivos denunciados o descubiertos con el fin de esclarecer los hechos y tomar medidas judiciales con los autores de dichos delitos.

Entre los sistemas informáticos que se utilizan para combatir el delito se encuentra El SAJO, que permite el registro, control y seguimiento de los actos delictivos ocurridos a nivel nacional que sean denunciados. FICHAJES es otros de los sistemas informáticos que se utiliza para almacenar las fichas de las personas que son de interés policial o que son juzgados por un delito cometido. Además, permite conocer los rasgos de las personas que se introducen en el

sistema. El SADEP, SAIP, CUBAFIS y CIRCULADO, son otros sistemas informáticos que permiten el enfrentamiento al delito en la provincia.

El SABIO, contiene la información de un grupo de sistemas y es el único sistema en Cuba que permite realizar un análisis de la información de los otros sistemas en la PNR.

El SISDED, que permite integrar la información sobre el conocimiento de la delincuencia en la comunidad y utilizar herramientas que den la posibilidad de controlar y direccionar su enfrentamiento en tiempo real.

Un datamart es una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica.(*Datamart*, s. f.). Para poder establecer patrones delictivos a partir de estas diversas fuentes de datos es necesario agrupar esta información en una base de datos unificada, en este caso un datamart de datos relacionados con el delito.

En la construcción de un datamart para la obtención de los patrones delictivos, se realiza un proceso de Extracción, Transformación y Carga (ETL) de todos los datos disponibles en las bases de datos relativas al delito. Este proceso permitirá integrar de manera eficiente los datos útiles, logrando así una visión única global que permite la realización un análisis de datos efectivo.

Para conseguir un correcto proceso de ETL es necesario analizar y entender en detalles las fuentes de datos. Por lo que la etapa inicial del ETL se realiza un perfilado y mapeo de datos, lo cual garantiza el movimiento y transformación exitosa de los datos desde su origen al datamart. Para mayor eficiencia y seguridad se emplea un almacenamiento de datos intermedio (staging área) que sirva para agrupar todos los datos antes de cargarlo al datamart.

En la actualidad los datos se encuentran dispersos en diferentes sistemas lo que trae demoras en el proceso de recopilación de los mismos. Además, la cantidad de información y el número de datos que se recogen por los oficiales es muy amplia, lo que imposibilita la determinación de los patrones delictivos.

Como resultado de lo anteriormente expuesto se identificó el siguiente **Problema:** La realización de un proceso de Extracción, Transformación y

Carga (ETL) de las diversas fuentes de datos del delito para la creación de un datamart en el MININT de Guantánamo.

Objeto: El proceso de Extracción, Transformación y Carga (ETL) de las diversas fuentes de datos del delito en el MININT de Guantánamo.

Objetivo: Realizar el mapeo y perfilado de los datos del proceso de ETL de las diversas fuentes de datos del delito en el MININT de Guantánamo.

Objetivo: Realizar la etapa inicial del proceso de ETL de las diversas fuentes de datos del delito en el MININT de Guantánamo.

El **campo de acción:** El mapeo y perfilado de los datos del proceso de ETL de las diversas fuentes de datos del delito.
Esta investigación tiene como **idea a defender:** El mapeo y perfilado de los datos contribuirá a la realización satisfactoria de proceso ETL dentro la creación del datamart para la determinación de los patrones delictivos en el MININT de Guantánamo.
Para darle cumplimiento al objetivo, se propusieron las siguientes **tareas:**

1. Caracterizar el proceso de proceso de ETL de las diversas fuentes de datos del delito.
2. Estudiar las tecnologías que existen para la realización del ETL
3. Caracterizar las fuentes de datos para la realización del ETL.
4. Describir el proceso de mapeo y perfilado de los datos del ETL.
5. Crear el perfil y mapa de los datos del ETL.
6. Diseñar el área almacenamiento temporal (stagin área)

Para desarrollar las tareas se utilizaron los siguientes métodos de investigación:
Métodos teóricos.

Métodos teóricos:

Análisis y síntesis: para analizar la documentación examinada en la elaboración de la fundamentación teórica del objeto de estudio y las tecnologías.
Modelación: para modelar el perfil y mapa de datos propuesto.

Histórico-Lógico: posibilitó consultar información de diferentes autores para obtener referencias de los antecedentes, la evolución y el desarrollo de los procesos de ETL.

Inducción y deducción: este método permitió identificar las diferentes deficiencias encontradas a partir de la investigación realizada, haciendo posible establecer problemas más generales y permitiendo llegar a conclusiones, las cuales fueron concretadas en la propuesta.

Métodos empíricos:

Análisis de documentos: Se han manejado fuentes documentales bibliográficas entre las que se incluyen manuales (del software para extraer información, SISDED, SAJO, SADEP, SAIP, FICHAJE, CUBAFIS, CIRCULADO), revistas científico-técnicas, libros ,tesis de maestría doctorales.

Observación (indirecta): aplicando la técnica de revisión documental, de los manuales de los software existentes y que serán utilizados para extraer la información como: SADEP, SISDED, SAJO, entre otros.

La investigación presentada está dividida en dos capítulos:

Capítulo 1. Fundamentación teórica del proceso de Extracción, Transformación y Carga (ETL)

En este capítulo se realiza caracterización del proceso de ETL, se expone una panorámica de las tendencias y tecnologías actuales que existen para la realización del ETL. Además, se realiza la descripción del proceso y se definen las fuentes de datos para la realización del ETL.

Capítulo 2. El mapeo y perfilado de los datos del proceso de ETL de las diversas fuentes de datos del delito

En el presente capítulo se presentan los elementos teóricos de las dos tareas de la primera etapa del proceso de ETL: el perfilado y el mapeo de datos. Se presenta además, el perfil de todos los datos de las BD del MININT que se emplearán como fuente de datos para la realización de un Datamart.

Se realiza un mapeo de los datos desde las fuentes de datos a un almacenamiento temporal para facilitar su migración y se presenta el diseño de la BD del mismo.

Capítulo 1. Fundamentación teórica del proceso de Extracción, Transformación y Carga (ETL)

En este capítulo se realiza caracterización del proceso de ETL, se expone una panorámica de las tendencias y tecnologías actuales que existen para la realización del ETL. Además, se realiza la descripción del proceso y se definen las fuentes de datos para la realización del ETL.

1.1 Caracterizar el proceso de ETL de las diversas fuentes de datos del delito.

A principios de la década de los sesenta, el software de acceso a datos consistía en aplicaciones independientes, basadas en ficheros maestros almacenados en cinta magnética; lo que significaba un acceso secuencial a los datos. La aparición de los discos magnéticos en la década de los setenta representó un cambio cualitativo, éstos permitían el acceso directo a los datos (DASD, del inglés Direct Access Strorage Device), favoreciendo el desarrollo de nuevas organizaciones de ficheros. A partir de ese momento se produjo una acelerada evolución en la tecnología de acceso a datos que no ha parado hasta nuestros días.

ETL cobró popularidad en la década de 1970 cuando las organizaciones comenzaron a utilizar múltiples repositorios de datos, o bases de datos, para almacenar diferentes tipos de información de negocios. La necesidad de integrar datos que se diseminaran por estas bases de datos creció con rapidez. ETL se convirtió en el método estándar para extraer datos de diferentes fuentes y transformarlos antes de cargarlos en una fuente pretendida o destino.

A fines de la década de 1980 y principios de la década de 1990, surgen los primeros almacenes de datos (data warehouses) y los data marts. En sus origines fue definido como “un conjunto de datos orientado a temas, integrado, no volátil, variante en el tiempo, como soporte para la toma de decisiones”(*5.2.1 ALMACENES DE DATOS (DATA WAREHOUSE) - Maria del Socorro Rosas Gaspar*, s. f.; Curto, 2006)

Un almacén de datos actúa como un depósito central de información que se origina en una o más fuentes de datos. Los datos fluyen desde los sistemas

transaccionales y otras bases de datos relacionales al almacén de datos y generalmente consisten en datos estructurados, semiestructurados y no estructurados. Estos datos se cargan, procesan y consumen regularmente.(Almeida, 2017)

Un data mart (DM) puede verse como un pequeño almacén de datos, que cubre un área temática determinada y ofrece información más detallada sobre un departamento en cuestión.(Almeida, 2017). Los data mart y los almacenes de datos proporcionan acceso a datos para análisis complejos, revelación de conocimientos y toma de decisiones.

En esos años los almacenes de datos y data mart proveían acceso integrado a datos de múltiples sistemas: computadoras mainframe, minicomputadoras, computadoras personales y hojas de cálculo. Con el tiempo, el número de formatos, fuentes y sistemas de datos ha aumentado enormemente. Extraer, transformar, cargar (ETL) se convirtió en un proceso complejo y muy importante para la creación de almacenes de datos y data mart.(*Pentaho Kettle Solutions*, s. f.), definen ETL como un conjunto de procesos para llevar los datos de los sistemas de origen a un almacén de datos o data mart.

Con el desarrollo de estas tecnologías diferentes departamentos eligieron diferentes herramientas ETL para utilizarlas con almacenes de datos distintos. Junto con fusiones y adquisiciones, muchas organizaciones terminaban con diferentes soluciones ETL que no estaban integradas. A continuación se muestra una figura que muestra el flujo de trabajo típico para la creación de un almacén de datos o data mart.

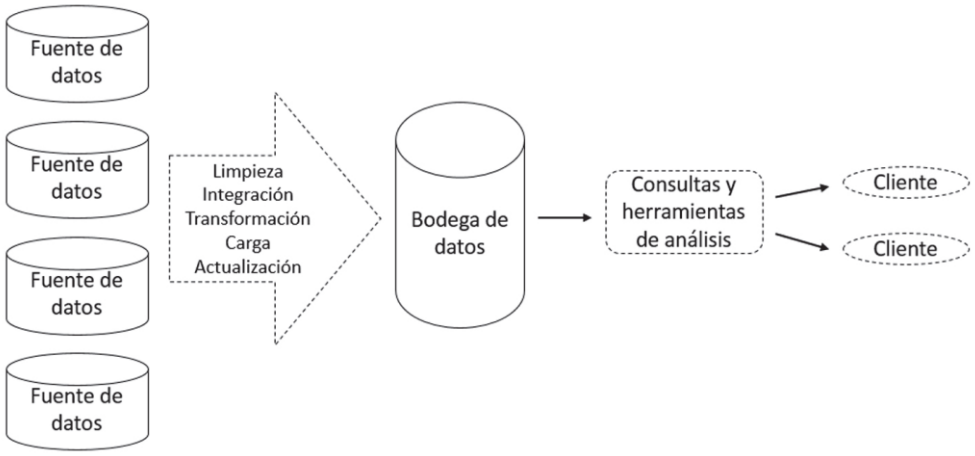


Figura 1.1 Marco de trabajo típico para la construcción de un almacén de datos (Han et al., 2011) (Inc, s. f.)

1.1.1 Preparación de los datos

Muchas de las cuestiones que rodean a los sistemas de apoyo para la toma de decisiones, se refieren en primer lugar a las tareas de obtener y preparar los datos. Los datos deben ser extraídos de diversas fuentes, limpiados, transformados y consolidados en la base de datos de apoyo para la toma de decisiones. Posteriormente, debe ser actualizado periódicamente. Cada una de estas operaciones involucra sus propias consideraciones especiales.

Integración de datos (ETL)

Es el proceso *Extraer, transformar y cargar* que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos, data marts o data warehouse para analizar, o en otro sistema operacional para apoyar un proceso de negocio.

Primera etapas en el ETL

Los procesos ETL pueden ser muy complejos. Un sistema ETL mal diseñado puede provocar importantes problemas operativos. Es por esto que la primera etapa del ETL se debe realizarse el perfilado y mapeo de los datos para garantizar que este se realice de manera correcta.

En un sistema operacional el rango de valores de los datos o la calidad de éstos pueden no coincidir con las expectativas de los diseñadores a la hora de especificarse las reglas de validación o transformación. Es recomendable realizar un examen completo de la validez de los datos (**perfilado de datos**) del sistema de origen durante el análisis para identificar las condiciones necesarias para que los datos puedan ser tratados adecuadamente por las reglas de transformación especificadas. Esto conducirá a una modificación de las reglas de validación implementadas en el proceso ETL.

El **mapeo de datos** es el proceso por el que se establecen correspondencias entre campos de una base de datos a otra. Es el primer paso para facilitar la migración, integración y otras tareas de gestión de datos. Antes de que los datos puedan ser analizados para obtener información de utilidad, deben homogeneizarse para que resulten accesibles a aquellos que tomen decisiones Ahora los datos llegan de muchas fuentes distintas y cada una puede definir puntos de datos semejantes de distinta forma. Por ejemplo, el campo Provincia de un sistema puente puede indicar Guantánamo como “Guantánamo”, pero otro sistema distinto puede que lo almacene como “GTMO” en la base datos destino deberá definir en qué campo y de qué manera se almacenará esta información.

El **mapeo de datos** salva la distancia entre dos sistemas, o modelos de datos, de modo que, cuando los datos se trasladen desde una fuente, lleguen veraces y útiles a su destino.

Extracción

(Teiken, 2012) y (*Pentaho Kettle Solutions*, s. f.) describen la etapa de extracción como un paso eso implica todo el procesamiento que se requiere para conectarse a varias fuentes de datos, extraiga los datos de estas fuentes de datos, y pone los datos a disposición del procesamiento posterior pasos. El proceso de extracción tiende a ser intensivo en entradas y salidas por lo tanto, puede interferir con las operaciones críticas.

Una vez que los datos se extraen van para un área temporal (staging area) y allí pueden ser limpiados. La limpieza es considerada como una de las más pasos importantes en el proceso ETL, ya que garantiza la calidad de la datos en el almacén de datos antes de aplicar la transformación.

Limpieza de datos

Pocas fuentes de datos controlan adecuadamente la calidad de los datos. Los datos requieren frecuentemente de una limpieza antes de que puedan ser introducidos. Las operaciones de limpieza típicas incluyen:

- El llenado de valores ausentes, la corrección de errores tipográficos y otros de captura de datos.
- El establecimiento de abreviaturas y formatos estándares.
- El reemplazo de sinónimos por identificadores estándares, etcétera.
- La validación de caracteres
- La validación de direcciones, códigos postales, url, email, etc.

Los datos que son erróneos y que no pueden ser limpiados, serán reemplazados. La información obtenida durante el proceso de limpieza puede ser usada para identificar la causa de los errores en el origen y por tanto, mejorar la calidad de datos.

Transformación

La transformación implica cualquier función aplicada al extraído datos entre la extracción de las fuentes y la carga de datos en objetivos (*Pentaho Kettle Solutions*, s. f.). Una vez que los datos se limpian, la transformación se puede aplicar en forma de conversión, generando agregados y claves sustitutas, ordenando y derivando calculados medidas. Las reglas comerciales complejas también se pueden aplicar en este paso.

El proceso de transformación se debe realizar cada vez que se realice una extracción de datos. Hay que destacar que la transformación de los datos permite establecer el nivel de granularidad3 que se tendrá finalmente.(Inmon, 2002)

Carga de datos

La carga de datos consiste en la incorporación de estos al almacén de datos o data mart con el formato adecuado. Se debe comprobar si los datos subidos coinciden con los datos procedentes de la transformación realizada.

Se trata generalmente de un paso sencillo, pero es muy crítico, puesto que si la información cargada no es la deseada o se produce algún error durante el proceso de carga. Los datos, en este caso, pueden llevar a resultados equivocados y por ende a la toma de decisiones errónea.

En este paso se puede decidir la información que se desea cargar, por lo que se puede establecer, según las transformaciones realizadas, el nivel de granularidad o detalle de la información que tendremos en el almacén de datos o data mart. Sin embargo, la situación más adecuada sería la elección del nivel de granularidad en la etapa de transformación de los datos y realizar la carga completa de estos, ya que así, se pueden comprobar de una forma más sencilla, si los datos cargados son o no los correctos y si se produjo un error.

Al igual que el proceso de extracción y de transformación, este proceso hay que realizarlo cada vez que hay una actualización de los datos. (Brachman & Anand, 1996; *Descubrimiento Patrones Desempeño Académico - Libro | Procesamiento de datos*, s. f.)

1.1.2 Técnicas para modelar el proceso de ETL

El modelado del proceso de ETL, al igual que el de cualquier objeto computacional, puede representarse utilizando tres niveles de abstracción: conceptual, lógico y físico. En la siguiente tabla se puede apreciar una breve descripción y comparación de los mismos.

Tipos	**Niveles de detalle**	**Objeto y conceptos**	**Ejemplo**
Conceptual	Bajo	Fuentes, atributos, transformaciones	La información de los clientes, está en el sistema transaccional; debe calcularse la edad de los

		y estructura de destino.	clientes previo a su inserción en el almacén de datos.
Lógico	Medio	Tablas de origen, dimensiones, tablas de hechos, Atributos, operaciones (aritméticas, lógicas y etc.)	La información del cliente debe calcularse así: extraer el año de nacimiento de la tabla ABC y retárselo al actual). Para luego insertarlo en el atributo edad de la dimensión cliente.
Físico	Alto	Tablas de origen, dimensiones, tablas de hechos Atributos, tipos de datos, precisión, restricciones, índices, entre muchas más.	La edad es representada por un entero de un byte, y para poder restársela al año actual, el valor extraído de ABC debe transformarse a tipo entero de un byte, luego restarlo al año que se extraerá de la fecha del sistema, posteriormente el valor se insertará en el atributo edad de la dimensión cliente, el cual está indexado.

Tabla 1.1 Tipos de modelado de ETL (Martínez et al., s. f.)

1.2 Estudiar las tecnologías que existen para la realización del ETL

Los procesos ETL suelen ser muy complejos, sobre todo por la gran cantidad de información que existe en la actualidad, que proviene de diferentes fuentes de información con diferentes estructuras y que se intentan integrar en un entorno homogéneo.

Existen varios tipos de herramientas especializadas para el ETL, que se diferencian según el formato en el que se encuentran los datos, el objetivo

perseguido, y tecnología utilizada. Algunas de las herramientas que se pueden encontrar fácilmente en la actualidad se presentan en la siguiente tabla.

Nombre	**Descripción**	**URL**
Informática Power Center	Es una plataforma de integración de datos empresariales que funciona como unidad para intercambio de datos, integración de datos en la nube, migración de datos, procesamiento de eventos complejos, enmascaramiento de datos, calidad de datos, replicación y sincronización de datos, virtualización de datos, gestión de datos maestros, y mensajería.(*Validación de Técnicas de Migración y Herramientas Etcl* | *Servidor SQL de Microsoft* | *SQL* | *Prueba gratuita de 30 días* | *Scribd*, s. f.)	**https://www.informatica.com/co/products/data-integration/powercenter.html** Con licencia
IBM Infosphere DataStage	Utiliza las características de un framework en paralelo de alto rendimiento y la notación gráfica para integrar datos en múltiples sistemas. Proporciona una potente plataforma escalable para la integración fácil y flexible de todo tipo de datos, incluidos big data en reposo (basado en Hadoop) o en movimiento (basado en secuencias), en plataformas distribuidas y *mainframe*. Gestiona la carga de trabajo y las reglas de	**https://www.ibm.com/us-en/marketplace/datastage** Con licencia

	negocio mediante la optimización del hardware. Está disponible en varias versiones, como Server Edition, Enterprise Edition y MVS Edition. Enterprise Edition presenta arquitectura de procesamiento paralela y trabajos paralelos. La edición de servidor representa principalmente los trabajos de servidor. La Edición MVS relacionada con trabajos de *mainframe*.(*Herramientas Big Data más usadas en la actualidad*, s. f.)	
Oracle Data Integrator (ODI)	Es una aplicación de software basada en ETL, que se utiliza para la transformación y fusión de datos o la integración de datos de alto volumen, alto rendimiento, hasta procesos basados en eventos y servicios de datos habilitados para SOA4 mediante el agregado de paralelismo. El componente de arquitectura importante de ODI es el repositorio, que es la recopilación de todos los metadatos y se accede mediante el modo cliente-servidor o el modo de cliente ligero. Oracle Data Integrator también funciona en el área de preparación y transformación como soporte para otro software de Oracle.(▷ 【 *Oracle ODI* 】 *Información, Reseñas y Precios* |	**http://www.oracle.com/technetwork/middleware/data-integrator/overview/index.html** Con licencia

	2020 |, s. f.)	
Microsoft SQL Server Integration Services (SSIS)	Es un componente de la base de datos SQL Server que realiza la integración de datos en el entorno de Windows. La principal ventaja de SSIS es que no es costoso. Sin embargo, una desventaja significativa es que no funciona en un entorno que no sea Windows. SSIS se lanzó por primera vez con SQL Server 2005. SQL Server 2008, 2012 también ha enriquecido el servicio de integración. En junio de 2016, se lanzó una nueva versión de SSIS. (*El paquete de SSIS*, s. f.)	**https://docs.microsoft.com/en-us/sql/integration-services/sql-server-integration-services?view=sql-server-2017** Con licencia
SAS ETL Studio	Ofrece una plataforma ETL integrada. SAS es uno de los líderes del mercado que combina aplicaciones de almacenamiento de datos e inteligencia para el proceso comercial tradicional. Proporciona la facilidad de extracción de datos multiproceso para acelerar la transferencia de datos y las operaciones relacionadas. SAS ayuda a reducir los datos duplicados o inexactos al proporcionar una interfaz de arrastrar y soltar, no necesaria de programación o SQL (lenguaje de consulta estructurado) para gestionar datos. SAS Data Integration Studio permite a los	**https://www.sas.com/en_us/software/data-management.html** Con licencia

	usuarios crear y editar rápidamente la integración de datos, capturar y gestionar automáticamente metadatos estandarizados desde cualquier fuente, visualizar y comprender fácilmente los metadatos empresariales. (*Did you know? Sabía que SAS es número 1 en inteligencia artificial y analítica* | *SAS*, s. f.)	
SAP Data Manager	SAP ha desarrollado un producto ETL con fuerte soporte para Hadoop5, transmisión de datos y aprendizaje automático, que permite integrar grandes cantidades de información de forma sencilla.	**https://www.sap.com/latinamerica/products/data-services.html** Con licencia
Pentaho Data Integration	Integración de datos utilizando un enfoque basado en metadatos. Utiliza un entorno gráfico intuitivo. No hace falta escribir líneas de código para su utilización y dispone de plugins.	**http://community.pentaho.com/projects/data-integration/** Con licencia y versión gratis
Talend Data Integration	Herramienta basada en Eclipse, para el proceso ETL que es uno de los procesos más importantes en la integración de datos.	**https://es.talend.com/products/talend-open-studio/** Con licencia y versión gratis
OpenRefine	Es una poderosa herramienta para trabajar con datos desordenado, limpiándolos, y transformándolos a	**http://openrefine.org/**

	un formato deseado.	Versión gratis
Scriptella ETL Project	Herramienta de lanzamiento de script ETL. Utiliza sintaxis XML para sus scripts, los cuales pueden integrarse con scripts escritos en SQL, JavaScrot, JEXL, Velocity, etc. Algunas de las fuentes de entrada que acepta son LDAP, JDBC, XML, CSV, texto, entre otros.	**http://scriptella.javaforge.com/** Versión gratis
Together	Se compone de varias herramientas separadas con funcionalidades ETL. Están desarrolladas en código Java y soporta la conexión con diferentes tipos de bases de datos (MSSQL, Oracle, DB2, QED, JDBC, MySQL,…) y acepta como entrada varios tipos de archivos (CSV, XML,…). Algunas herramientas son: TDC – Together Document Converter, TDT – Together Data Transformer, TXE – Together XML Extractor.	**http://www.together.at/download** Versión gratis
Xineo XIL	Define un lenguaje XML para transformar fuentes de datos basadas en registros en archivos XML. Soporta JDBC y estructuras de texto.	**http://software.xineo.net/xil.jspx** Versión gratis
CloverETL Community Edition	Es una herramienta muy gráfica que permite varios tipos de transformaciones, así como diversos	**http://www.cloveretl.com/products/community-edition**

	tipos de entrada y salida de datos, como son los procedentes de las BBDD MySQL, PostgreSQL, SQLite, MSSQL, Oracle, Sysbase y Derby, archivos CSV, XML, etc. Cuenta con versiones de pago que permiten muchas más opciones (clasificación, clusters).	Versión gratis
Apatar	Usa interfaz gráfica de trabajo mediante la cual se puede hacer el filtrado, la validación y la planificación de los datos. Los conectores incluyen MySQL, PostgreSQL, Oracle, MSSQL, Sybase, FTP, HTTP, SalesForce.com, SugarCRM, Compiere ERP, CRM Goldmine, XML, archivos planos, WebDAV, Buzzsaw, LDAP, Amazon y Flickr. No se requiere. Todos los metadatos se guardan en archivos XML.	**http://www.apatar.com/** Versión gratis
Jaspersoft ETL	Herramienta basada en Eclipse, para el proceso ETL que es uno de los procesos más importantes en la integración de datos. Incluye flujos y procesa diferentes tipos de archivos. Fácil de desplegar.	**http://community.jaspersoft.com/project/jaspersoft-etl** Versión gratis
Data Pipeline	Transforma datos y los procesa. Puede leer y escribir archivos de tipo CSV, Excel, JDBC, JSON.	**http://northconcepts.com/data-pipeline/** Versión gratis

KETL	Está basado en java. Incluye gestión de Jobs y alertas. Es capaz de gestionar varios hilos a la vez. Los Jobs están definidos en XML.	**http://www.ketl.org/** Versión gratis

Tabla 1.2 Herramientas ETL empresariales y de código abierto (*Principales categorías de herramientas ETL - BI Geek Blog*, s. f.)

Se puede observar que existen varias herramientas ETL, que pueden facilitar los procesos de extracción, transformación y carga de datos. Las más interesantes son las herramientas que utilizan interfaces gráficas (GUI), ya que son más fáciles e intuitivas de utilizar.

Las herramientas mencionadas, usan técnicas muy diferentes a la hora de manipular los datos, utilizan distintos tipos de elementos de entrada y salida, y al ser varias herramientas de código abierto, algunas fueron desarrolladas por necesidades propias del programador. Por todo ello hay que tener claro, los tipos de datos, de dónde serán extraídos y qué formato deben de tener estos para poder usarse posteriormente en el resto de componentes BI.

Todas las herramientas de ETL citadas anteriormente son buenas opciones, pero dependiendo de las necesidades, el entorno de trabajo, las posibilidades o la estrategia de negocio, algunas pueden encajar mejor que otras en alguna tarea. Dependiendo de las motivaciones, se pueden establecer preferencias.

De las herramientas anteriores existen tres que son las que presentan las mejores capacidades de extraer, transformar y cargar datos, que hacen que sean de gran utilidad:

- Pentaho Data Integration
- Talend Data Integration
- OpenRefine

Para su selección se tuvo en cuenta, de qué fuentes las herramientas pueden extraer datos (bases de datos, la nube, archivos planos, archivos de Excel, XML, big data, entre otros) y la facilidad con que lo hacen, siendo este un punto

en consideración ya que si una herramienta que permita poder extraer datos de estas distintas fuentes será de gran utilidad para el cumplimiento del objetivo buscado, otra consideración es la transformación de datos, es decir (filtrar, unir, limpiar, entre otros) y la facilidad con que lo hacen, permitiendo procesar toda la información correspondiente del proceso de extracción, y por último está el proceso de carga de datos, en el cual se tuvo en cuenta, las distintas fuentes en las que se puede cargar los datos es decir (bases de datos, la nube, archivos planos, archivos de Excel, XML, big data, entre otros) y la facilidad con que lo hacen.

1.2.1 Herramientas para el perfilado de datos

La creación de perfiles de datos, una actividad tediosa e intensiva en mano de obra, puede automatizarse con herramientas para hacer más factibles los grandes proyectos de datos. Estos son esenciales para su pila de análisis de datos.

Herramientas de creación de perfiles de datos de código abierto.

1. Quadient DataCleaner(*DataCleaner*, s. f.)

   Las características fundamentales son:

   - Calidad de datos, perfiles de datos y disputas de datos
   - Detecta y combina duplicados
   - Análisis booleano
   - Análisis de integridad
   - Distribución de juego de caracteres
   - Análisis de brecha de fecha
   - Coincidencia de datos de referencia

2. Aggregate Profiler calidad y perfil de datos de código abierto (*Aggregate Profiler User Interface | Download Scientific Diagram*, s. f.) Las características fundamentales son:

   - Perfiles de datos, filtrado y gobierno
   - Comprobaciones de similitud

- Enriquecimiento de datos
- Alerta en tiempo real para problemas o cambios de datos
- Análisis de cesta con validación de gráfico de burbujas
- Vista única del cliente
- Creación de datos ficticios
- Descubrimiento de metadatos
- Herramienta de descubrimiento de anomalías y limpieza de datos
- Integración Hadoop

3. Talend Open Studio: un conjunto de herramientas de código abierto.(*ETL de código abierto e integración de datos gratuita: Talend Open Studio*, s. f.)

   Las características fundamentales son:

   - Evaluación de datos personalizables
   - Una biblioteca de patrones
   - Analítica con cuadros gráficos
   - Detección de patrones de fraude
   - Análisis de conjunto de columnas
   - Emparejamiento avanzado
   - Correlación de columna de tiempo

Herramientas de perfilado de datos comerciales

1. Data Profiling in Informática: (*Data profiling, el primer paso en calidad de datos*, s. f.; *What Is Data Profiling and How Does It Make Big Data Easier?*, s. f.)

   Las características fundamentales son:

   - Consola de administración de datos que imita el flujo de trabajo de administración de datos
   - Interfaz de manejo de excepciones para usuarios comerciales
   - Gobierno de datos empresariales
   - Mapee las reglas de calidad de datos una vez e impleméntelas en cualquier plataforma
   - Estandarización de datos, enriquecimiento, desduplicación y consolidación.

- Gestión de metadatos

2. Oracle Enterprise Data Quality.

Las características fundamentales son:

- Perfiles de datos, auditorías y paneles de control.
- Análisis y estandarización, incluidos campos construidos, datos archivados incorrectamente, datos mal estructurados y campos de notas
- Combinación y combinación automatizadas.
- Gestión de casos por operadores humanos.
- Verificación de dirección
- Verificación de datos del producto
- Integración con Oracle Master Data Management

3. SAS DataFlux.(*SAS Help Center: About SAS and DataFlux*, s. f.)

Las características fundamentales son:

- Extrae, limpia, transforma, conforma, agrega, carga y gestiona datos
- Admite la gestión de datos maestros orientada a lotes y en tiempo real
- Crea servicios de integración de datos reutilizables en tiempo real.
- Capa de datos de referencia semántica fácil de usar
- Visibilidad de dónde se originaron los datos y cómo se transformaron
- Componentes opcionales de enriquecimiento

1.2.2 Herramientas de creación de mapeo de datos

Las herramientas de mapeo de datos permiten a los desarrolladores definir las reglas de mapeo a través de la codificación. La mayoría de las herramientas proporcionan una interfaz gráfica también para definir estas reglas de mapeo y esto a su vez facilita a las personas no técnicas definir las reglas de mapeo.

Herramientas de creación de mapeo de datos de código abierto.

1. **Talend** (*ETL de código abierto e integración de datos gratuita: Talend Open Studio*, s. f.)

Las características fundamentales son:

- Proporciona más de 900 componentes prefabricados.
- Integración perfecta con su entorno.

- Es escalable según sus datos.

2. **Informática**

Las características fundamentales son:

- Plataforma de integración de datos ágil totalmente integrada.
- Integración con Power Center.
- Los conectores proporcionarán conectividad de alto rendimiento a los datos.
- Puede realizar el intercambio de datos B2B.

3. **Altova** (*Soluciones XML, de integración de datos y de desarrollo móvil de Altova*, s. f.)

Las características fundamentales son:

- Altova puede realizar cualquier mapeo para cualquier tipo de datos como XML, JSON, datos de bases de datos, archivos de texto y planos, XBRL, EDI, SOAP y REST, servicios web, Excel, Buffers de protocolo de Google.
- Proporciona una interfaz gráfica para mapear, visualizar, manipular y ejecutar proyectos de mapeo individuales y complejos.
- Tiene una biblioteca extensible de procesamiento de datos y funciones de conversión.
- Le permitirá importar el código de transformación de datos existente.

Herramientas de creación de mapeo de datos comerciales.

1. **CloverDX** (*CloverDX | Solve demanding, real-world data challenges*, s. f.)

Las características fundamentales son:

- Adecuado para realizar tareas simples y complejas.
- Puede diseñar transformaciones de datos reutilizables.
- Se puede conectar con los sistemas externos a través de API, colas de mensajes, observadores de archivos y activadores de eventos.

- Le permitirá programar, administrar y monitorear flujos de trabajo complejos.
- Se puede gestionar cualquier número de trabajos.

2. **Pentaho** (*¿Qué es Pentaho Data Integration (PDI) y para qué sirve?*, s. f.)

Las características fundamentales son:

- Con Pentaho, podrá cambiar sin problemas entre los motores de ejecución como Apache Spark y Pentaho.
- Proporciona soporte robusto para distribuciones de Hadoop, Spark, NoSQL y almacenes de objetos.
- Supervisión del rendimiento.
- Revertir trabajo y reiniciar.

3. **Salesforce** (*CRM On Demand, Soluciones CRM On Demand de Salesforce - Salesforce España*, s. f.)

Las características fundamentales son:

- Le ayudará a conectar cualquier fuente de datos.
- Conjunto amplio de API.
- Los sistemas de back-office también se pueden conectar.

1.3 Caracterizar las fuentes de datos para la realización del ETL.

1.3.1 Descripción de las diversas fuentes de datos del delito.

Cuando se procede a realizar la búsqueda de una información de un hecho delictivo o hacer un análisis respecto a una información de una denuncia, el oficial debe acceder a diversas fuentes de datos de sistemas informáticos existentes en la PNR.

Verificar denuncia

Si el oficial desea verificar una denuncia, la misma es registrada en un módulo en el sistema informático SISDED, el cual brinda información obtenida desde la base de datos del sistema informático SAJO, donde se almacenan las denuncias que se realizan en las estaciones de la PNR, se obtiene: fecha,

lugar, consejo popular en que ocurrió el hecho, tipo de delito y si el caso ha sido esclarecido o no, proporcionando también el número de autores que cometieron el delito. También recibe una descripción del caso, si una persona ha sido procesada, y si criminalística realizó levantamiento de huellas.

Búsqueda de PIP

Si en la investigación que se realiza por el oficial, se precisa conocer las personas que son de interés policial (PIP), para investigar a quienes se consideren como posibles autores. El oficial accede al sistema SISDED, donde se encuentra el Módulo de Interés Policial, para obtener el listado de estos PIP. En este módulo se visualizan el potencial delictivo por categorías:

- PIP: personas que la policía controla por su mal comportamiento en la comunidad y son proclive a cometer delito.
- PIPP: personas que la policía controla por su mal comportamiento en la comunidad, que son potencial delictivo y a las que se les abre un expediente para controlarlos.
- PIPP 201: ex reclutas que cometieron delitos grabes y salieron en libertad, mantienen mala conducta y se le mantiene un expediente abierto para su control.
- INT_201: ex reclutas que salieron en libertad y mantienen una buena conducta, no se controlan, pero se mantiene un registro de ellos.
- PIP Indultado: ex reclutas que salieron en libertad por un indulto, pero no mantienen una buena conducta y son PIP.

El oficial obtiene la lista de PIP en tres maneras, por municipio, por estación que solicitó el interés de esa persona y por consejo popular. Con esta información se solicitan y entrevistan en la estación de policía los PIP que considera el oficial de la policía; entrevista que posibilita descartar a posibles autores, que brinda informaciones que permiten el desarrollo de la investigación y el esclarecimiento del caso.

Comparando evidencias

Durante la toma de datos en el lugar del hecho con la PNR participa la Criminalística, la cual se encarga de levantar huellas de olor y dactilares,

además de fotos y otras evidencias. Se analizan los datos recogidos y si ellos arrojan algún resultado son enviadas al investigador de la PNR en formato plano, además de que son insertadas en el sistema que utiliza la Criminalística (Cubafis). Las huellas dactilares que se recogen se cotejan con la base de datos Cubafis y si ya la persona dueña de las huellas se encuentra en este sistema, a lo que se denomina positivo Cubafis, se podrá obtener los datos de dicha persona y tras investigación se puede declarar como autor del hecho.

Búsqueda por características de la persona

El oficial, se entrevista con personas que fueron testigos del hecho o al menos tienen un mínimo de conocimiento de lo ocurrido, haciendo preguntas que le posibiliten tener una impresión inicial del caso. Durante el cuestionario que realiza el oficial pueden salir a luz características físicas o rasgos distintivos de la persona o grupo de personas que cometieron el delito o infracción. Con esta información el oficial consulta el sistema de Fichajes, realizando un filtro o búsqueda de personas que coincidan con las características que tomó de las entrevistas, teniendo como resultado una lista de nombres de personas que coincidan con dichas características, y con algún otro conocimiento referente al caso, realizar otro filtro reduciendo aún más la lista de nombres que obtuvo con la primera búsqueda.

Descarte de reclusos y detenidos

En la investigación realizada por el oficial encargado de la denuncia, pueden aparecer nombres de personas que ya ha sido juzgada por la justicia por haber cometido algún otro delito. El oficial se apoya en el sistema informático SADEP, sistema que lleva el control de los reclusos, donde conocer si alguno de los sospechosos o posibles autores se encuentra en la cárcel cumpliendo condena, y si es el caso, saber si estuvo de pase para la fecha en que ocurrió el hecho. También se conoce si el posible autor ha sido puesto en libertad, a través de una condicional o por el cumplimiento de la condena impuesta. Además, el oficial recibe información acerca de los reclusos que están vinculados a algún empleo y donde, y los reclusos que luego de salir de pase no volvieron al lugar donde están cumpliendo la condena, quedando en el sistema como que están prófugos.

El oficial; además del SADEP consulta el sistema informático SAIP, con el objetivo de saber si alguno de los nombres de los posibles autores que surgieron durante la investigación se encuentra en este sistema. El encontrar el nombre de un posible autor indica que esta persona estuvo detenida en alguna de las estaciones policiales, posibilitando el descarte de esta persona como un posible autor si la fecha y el tiempo de detención coincide con el de la ocurrencia del hecho.

1.3.2 Descripción de los sistemas automatizados que existen.

Sistema de Dirección para el Enfrentamiento al Delito (SISDED).

Este sistema permite integrar la información sobre el conocimiento de la delincuencia en la comunidad e implementar herramientas que posibiliten (en los diferentes escalones de mando de la policía, desde los mandos provinciales hasta el de jefe de sector) controlar y direccionar su enfrentamiento en tiempo real, brindando facilidades para la rapidez operativa y toma de decisiones. Además de brindar información a dichos órganos para mejorar el esclarecimiento y/o continuidad de su trabajo según el caso en cuestión.

El SISDED posee gran potencial informativo sobre personas, pues utiliza la información registrada en las bases de datos del MININT, logrando la búsqueda de sus antecedentes. Además, genera estadísticas desde lo más simple hasta lo más complejo en forma de tablas y gráficos para realizar análisis del comportamiento del delito y personas de interés, potenciando la toma de decisiones oportunas.

Analítico Provincial del Sistema Automatizado Jurídico Operativo (SAJO):

Tiene por objetivo esencial el registro, control y seguimiento de los hechos delictivos ocurridos en cualquier lugar del territorio nacional que sean denunciados; así como otros casos que no constituyen delitos como: los índices de peligrosidad, las muertes, lesiones y daños por accidentes de tránsito y los ausentes a domicilio.

Este sistema constituye el medio automatizado que permite medir el trabajo de las unidades de enfrentamiento, principalmente las del orden interior y el

comportamiento del delito en los territorios en diferentes períodos, apoya y facilita los análisis de la situación en los distintos niveles de Dirección.

Sistema Único de Fichaje Criminal:

Este sistema está compuesto por tres módulos principales: Fichaje, Búsqueda y Administración. Adicionalmente contiene una capa de servicios que posibilita la integración con otras aplicaciones. Se implementa además un Cuadro de Mando que permita reflejar las principales métricas del proceso de fichaje en el país. Esta arquitectura posibilita una integración adecuada con las pautas trazadas por el MININT para el despliegue e integración de aplicaciones dentro de su red privada.

Sistema Automatizado Dirección de Establecimientos Penitenciarios (SADEP)

Es un sistema que lleva el control de los reclusos. Además, cuenta con las informaciones referentes a la situación de esas personas. Posee el conocimiento de los reclusos que ya fueron dados en libertad, los que fueron libertados a raíz de la visita del Papa San Francisco a Cuba, es decir, los indultados y de los reclusos de pase, los empleados y los que están prófugos.

Sistema Automatizado de Información a la Población (SAIP):

Es utilizado mayormente en las estaciones de policías, para llevar el control de las detenciones realizadas. Este sistema cuenta con la información del detenido, causas de su detención, estación en la que se realizó dicha detención y el rango de tiempo (dado por días y hora) que fue detenido.

Sistema Cubafis:

Empleado por el órgano de la Criminalística, permite conocer si en un caso determinado luego de hacer la recogida de evidencias, se encuentra la detección de huellas dérmicas, y además posibilita saber si estas han cotejado con una persona del sistema o con un sospechoso, o si aún no son identificables.

Sistema de Circulación de Personas:

Lleva el control de las personas que son autores o posibles autores de un hecho. Personas que fueron citadas y nunca aparecieron a la hora que se le

citó ni tampoco se han presentado a la delegación para dar su testimonio en el rango de diez días luego de haber sido citados. Se puede decir que estas personas se han dado a la fuga y por tanto se le da entrada al sistema, lo que permite que esta persona pueda ser identificada en cualquier parte del país, pues es un sistema nacional, y además comparte esta información con empresas que brindan servicio a la población como las estaciones de viajes.

1.4 Deficiencias en el proceso

1. El MININT consta de diferentes sistemas informáticos que tratan el delito.
2. El volumen de informaciones que brindan los sistemas existentes es muy grande.
3. Los sistemas existentes, además de la información del delito, no facilitan una dirección a la investigación en cuestión, la cual debe ser determinada por el investigador.

1.5 Mejoras propuestas

- Se propone una base de datos que permita visualizar toda la información de interés referente al delito.
- Reducir el volumen de información que es analizada por el investigador, atendiendo a las características del delito en investigación.

Capítulo 2. El mapeo y perfilado de los datos del proceso de ETL de las diversas fuentes de datos del delito

En el presente capitulo se presentan los elementos teóricos de las dos tareas de la primera etapa del proceso de ETL: el perfilado y el mapeo de datos. Se presenta además, el perfil de todos los datos de las BD del MININT que se emplearán como fuente de datos para la realización de un Datamart.

Se realiza un mapeo de los datos desde las fuentes de datos a un almacenamiento temporal para facilitar su migración y se presenta el diseño de la BD del mismo.

2.1 Primera etapa del ETL

Para la realización de un exitoso proceso de ETL es necesario obtener toda la información posible de las fuentes de datos para poder realizar cada una de sus etapas. Por lo que todo proceso de ETL tiene como etapa inicial un perfilado y mapeo de los datos. Estas técnicas facilitaran la extracción y transformación de los datos para su posterior carga.

2.1.1 El perfilado de datos (data profilling)

El perfilado de datos es el proceso de recopilación de estadísticas y otra información sobre los datos existentes en nuestros orígenes de información. Esta información va a ser de gran utilidad para el diseño de los procesos ETL. También puede ser parte importante de cualquier iniciativa de calidad de datos, ya que antes de que la calidad de estos se pueda mejorar, habrá que establecer cuál es el estado actual de los datos, y para ellos podemos valernos de estas técnicas.
El perfilado de datos le permite responder las siguientes preguntas sobre los datos que surgirán a lo largo del proceso ETL:

- ¿Están completos los datos?
- ¿Hay valores en blanco o nulos?
- ¿Son únicos los datos?
- ¿Cuántos valores distintos hay?
- ¿Los datos están duplicados?
- ¿Hay patrones anómalos en sus datos?

- ¿Cuál es la distribución de patrones en sus datos?
- ¿Son estos los patrones que esperas?
- ¿Qué rango de valores existen y se esperan?
- ¿Cuáles son los máximos, mínimos, y valores promedio para datos dados? ¿Son estos los rangos que espera?

El perfilado puede ser desarrollado a diferentes niveles y técnicas según la necesidad existente, el acceso a los datos o la características de los mismos.
Los niveles más utilizados son:
Column profile: recopilación de estadísticas sobre los datos existentes en una columna individual.
Dependency profile: análisis entre las dependencias de las diferentes columnas de una tabla.
Join profile: chequeos de dependencias entre diferentes tablas.

El punto de partida para el perfilado puede ser el **Column Profile**, que nos puede proporcionar información tan interesante como:

- **Tipo de datos de las columnas:** permite conocer los tipos de datos que pueden variar según el gestor de base de datos que lo soporte.
- **Número de valores distintos**: cuantas entradas únicas contiene una determinada columna (en un análisis de clientes, que un valor exista varias veces puede, por ejemplo, reflejar registros duplicados).
- **Número de valores nulos (Null) o vacíos en la columna**: nos puede ayudar a identificar registros cuyos datos están incompletos.
- **Valores mínimos y máximos en el campo**, no solo a nivel numérico, sino también a nivel de texto.
- El uso de f**unciones estadísticas** como suma, mediana, media o desviación estándar puede ser útil también para sacar conclusiones sobre los datos.
- **Longitud de los campos y patrones de cadenas**: en muchas ocasiones los campos tendrán que tener un formato determinado que obligara a una determinada longitud o al uso de unos determinados patrones (como por ejemplo el carnet de identidad que tiene que tener solo 11 caracteres). El control de la longitud de los valores de la columna o la búsqueda de valores que no cumplan los patrones nos puede

ayudar a encontrar valores incorrectos. Con los patrones de cadenas y expresiones regulares podemos buscar determinadas ocurrencias de valores que determinen datos incorrectos igualmente.

- **Número de palabras, número de caracteres en mayúsculas y minúsculas**.
- **Contadores de frecuencia de ocurrencia de valores**.

Además de los análisis por columna, las herramientas se pueden descubrir **dependencias entre diferentes campos de una misma tabla**, como por ejemplo el análisis de los datos geográficos de un cliente y las relaciones entre los diferentes campos (codigo postal, población, provincia, región, etc).
El análisis de las relaciones entre tablas (**Join profile**) es más fácil de realizar, pues la verificación de dependencia entre valores de diferentes tablas puede ser sencilla de establecer determinando las tablas y sus relaciones con otras tablas a través de las llaves foráneas.

2.1.2 Mapeo de datos

El término "mapeo de datos" se refiere a la capacidad de hacer conexiones entre los campos de datos de origen y sus campos de datos de destino. Además, los técnicos y desarrolladores pueden crear conversiones de código para garantizar que se cumpla el resultado deseado. En todas las empresas y organizaciones actuales, el mapeo de datos preciso es una parte esencial de las estrategias de migración e integración de datos.
El mapeo de datos es lo que determina qué datos se transferirán a un determinado sistema de destino. Determina el formato de los datos que se transferirán, con qué frecuencia se realizarán estas transferencias y cómo se iniciarán o desencadenarán las transferencias de datos.
Este proceso permite la posterior transformación de cualquier campo de datos en el formato requerido, la longitud y el tipo de datos en un formato y que sean compatibles con un determinado receptor.
El mapeo de datos puede realizarse manualmente y esto es ideal para pequeñas organizaciones y empresas. Sin embargo, las empresas más grandes con redes y bases de datos más complejas, o aquellas que se basan

principalmente en la nube, se beneficiarían de las técnicas de mapeo de datos semiautomatizadas o automatizadas.
Al crear los mapas de datos se muestran las conexiones de datos entre los campos de origen y sus objetivos, el mapeo de datos permite a los realizadores del proceso ETL también prever problemas potenciales, como identificar campos o formatos de datos incompatibles. Dado que el mapa de datos señala esto antes de ponerlo en práctica, se pueden tomar medidas preventivas preventivas y de corrección.

El mapeo de datos es importante para mantener relaciones coherentes entre dos o más elementos de datos. Proporciona el medio común en el que todos los campos de datos de una organización interactúan entre sí. Como tal, el mapeo de datos es responsable de mantener la compatibilidad de datos en todo el procesamiento y análisis de datos.
El mapeo de datos también ayuda a identificar datos personales y privados y a mantenerlos seguros. Siempre que la asignación a los servidores que manejan este tipo de información sea precisa, los datos pueden cifrarse adecuadamente y transferirse de forma segura desde el origen al destino.
El mapeo de datos en las organizaciones se usa principalmente en dos tareas principales:

1. migración de datos
2. integración de datos.

En la **migración de datos**, la información se mueve de un campo de datos a otro. La función del mapeo de datos en la migración de datos es crear un esquema que muestre la ruta adecuada entre el campo de datos de origen y su destino. Un ejemplo simple de esto es cuando uno compra un nuevo teléfono inteligente Android. Después de hacer una copia de seguridad de los datos del teléfono anterior en la nube, el usuario ahora desea descargar esos datos (contactos, aplicaciones, archivos, fotos, configuraciones, etc.) al nuevo teléfono.
La **integración de datos** es el proceso de construcción y sincronización de datos entre dos o más fuentes. La integración de datos es responsable de crear un puente entre un modelo de datos antiguo y un nuevo modelo de datos. El mapeo de datos hace posible el fomento de esta conexión al garantizar que los

datos entre modelos sean compatibles y accesibles para ambos. El mapeo de datos es importante para rectificar las diferencias entre dos o más fuentes de datos y sus objetivos para garantizar la eficiencia de la integración.

En general, el mapeo de datos es un componente integral de la construcción y mantenimiento de datos, así como una forma segura de garantizar un exitoso proceso de ETL.

2.2 Perfilado y mapeo de datos de las fuentes de datos del MININT

2.2.1 El perfilado de los datos

Para el perfilado de datos se emplearán 2 niveles:

1. Perfilado de columnas: Caracterización de las columna de la fuente de datos
2. Perfilado de dependencia: Análisis entre las dependencias de las diferentes columnas de una tabla.

A continuación el perfilado de columnas obtenido por tablas de la fuentes de datos.

T_BPOLICIA

Campo	**Tipo**	**Nulo**
ADVERTENCIA	VARCHAR(50)	YES
C_IDENTIDAD	VARCHAR(11)	NO
COLOR_OJO	VARCHAR(20)	YES
CONTRAVENCION	VARCHAR(50)	YES
DESCRIPCION_ADVERTENCIA	VARCHAR(500)	YES
DESCRIPCON_MUNICIPIO	VARCHAR(20)	YES
ESTATURA	FLOAT	YES
ID_BPOLICIA	INTEGER	NO
IDENTIDAD	VARCHAR(30)	YES
MUNICIPIO	VARCHAR(50)	YES
PELO	VARCHAR(20)	YES
PIEL	VARCHAR(20)	YES
UNIDAD	INTEGER	YES

T_ELECT

Campo	Tipo	Nulo
APELL1	VARCHAR(30)	YES
APELL2	VARCHAR(30)	YES

C_IDENTIDAD	VARCHAR(11)	NO
CDR	VARCHAR(50)	YES
CIRCUNSCRIPCION	INTEGER	YES
COLEGIOS	VARCHAR(60)	YES
CONSEJOS	VARCHAR(20)	YES
DIRECCION	VARCHAR(50)	YES
FECHA_NACIMIENTO	DATE	YES
MUNICIPIO	VARCHAR(15)	YES
NOMBRE	VARCHAR(20)	YES
NOMBRE_CONSEJOS	VARCHAR(20)	YES
NOMBRE_MUNICIPIO	VARCHAR(15)	YES
NOMBRE_PROVINCIA	VARCHAR(20)	YES
PROVINCIA	VARCHAR(20)	YES
SEXO	VARCHAR(10)	YES
ZONA	VARCHAR(50)	YES

T_ARMAS DE FUEGO

Campo	Tipo	Nulo
C_IDENTIDAD	VARCHAR(11)	NO
CALIBRE	INTEGER	YES
CATEGORIA	VARCHAR(30)	YES
DEST_ARMA_OCUPACIN	VARCHAR(30)	YES
ESTADO_ ARMA	VARCHAR(40)	YES
ESTADO1	VARCHAR(40)	YES
FECHA_ALTA_ARMA	DATE	YES
FECHA_ESTADO	DATE	YES
FECHA_LICENCIA	DATE	YES
ID_ARMAS DE FUEGO	INTEGER	NO
MARCA	VARCHAR(30)	YES
MOTIVO_ESTADO	VARCHAR(50)	YES
PROCEDENCIA_ARMA	VARCHAR(40)	YES
PROVINCIA	VARCHAR(50)	YES
SERIE	INTEGER	YES
TIPO_ARMA	VARCHAR(50)	YES

T_CIRCULADOS

Campo	Tipo	Nulo
ANO_DENUNCIA	VARCHAR(50)	YES
C_IDENTIDAD	VARCHAR(11)	NO

CATEGORIA	VARCHAR(50)	YES
CIUDADANIA	VARCHAR(50)	YES
DELITO	VARCHAR(30)	YES
DIRECCION_CIRCULACION	VARCHAR(50)	YES
FECHA_CIRCULACION	DATE	YES
ID_CIRCULADOS	INTEGER	NO
MUNICIPIO_CIRCULADO	VARCHAR(40)	YES
NUMERO_DENUNCIA	INTEGER	YES
PROVINCIA	VARCHAR(30)	YES
PROVINCIA_CIRCULADA	VARCHAR(50)	YES
UNIDAD	VARCHAR(50)	YES

T_CUBAFIS

Campo	Tipo	Nulo
APELL1	VARCHAR(40)	YES
APELL2	VARCHAR(40)	YES
C_IDENTIDAD	VARCHAR(11)	NO
FECHA_ACTUAL	DATE	YES
FECHA_REGISTRO	DATE	YES
ID_CUBAFIS	VARCHAR(30)	NO
MUNICIPIO	VARCHAR(50)	YES
NOMBRE1	VARCHAR(20)	YES
NOMBRE2	VARCHAR(20)	YES
PROVINCIA	VARCHAR(50)	YES
SEXO	VARCHAR(50)	YES

T_ESTANCIA_CASERO

Campo	Tipo	Nulo
APELL1_HUESPED	VARCHAR(40)	YES
C_IDENTIDAD	VARCHAR(11)	NO
C_IDENTIDAD_CASERO	VARCHAR(11)	YES
CIUDADANIA	VARCHAR(50)	YES
DIRCCION_ESTANCIA	VARCHAR(50)	YES
FECHA_FIN	DATE	YES
FECHA_INICIO	DATE	YES
FECHA_NACIMIENTO_HUSPED	DATE	YES

ID_ESTANCIA_CASERO	VARCHAR(50)	NO
MUNICIPIO_ESTANCIA	VARCHAR(40)	YES
NOMBRE_HUSPED	VARCHAR(50)	YES
PASAPORTE	VARCHAR(50)	YES

T_EXPECOND

Campo	Tipo	Nulo	Campo
APELL1	VARCHAR(40)	YES	APELLIDOS
APELL2	VARCHAR(40)	YES	APELLIDOS
C_IDENTIDAD	VARCHAR(11)	NO	CARNET_IDENTIDAD
CHAPA	VARCHAR(15)	YES	MATRICULA
DESCRIPCION_LIMITANTE	VARCHAR(200)	YES	LIMITANTE
DESCRIPCION_TRAMO	VARCHAR(150)	YES	TRAMO
DESCRIPCION_UNIDAD	INTEGER	YES	NUMERO
ESTADO_LICENCIA	VARCHAR(50)	YES	ESTADO_LICENCIA
FECHA_EXP	DATE	YES	FECHA_EXPEDIDA
FECHA_VENC	DATE	YES	FECHA_VENCIMIENTO
FECHA_VENC_LICENCIA	DATE	YES	FECHA_VENCIMIENTO
ID_EXPECOND	INTEGER	NO	ID_EXPEDIENTE
LICENCIA	VARCHAR(20)	YES	LICENCIA
MUNICIPIO	VARCHAR(30)	YES	NOMBRE
NOMBRE	VARCHAR(30)	YES	NOMBRE
NOMBRE2	VARCHAR(20)	YES	NOMBRE
PROVINCIA	VARCHAR(20)	YES	NOMBRE

T_REGULADOS

Campo	Tipo	Nulo
C_IDENTIDAD	VARCHAR(11)	YES
CATEGORIA	VARCHAR(50)	YES
DESCRIPCION_CATEGORIA	VARCHAR(200)	YES
FECHA_ALTA	DATE	YES
ID_REGULADOS	VARCHAR(20)	YES
PROFESION	VARCHAR(30)	YES
REGULADOS_VITALES	VARCHAR(10)	YES

T_REGVEH

Campo	Tipo	Nulo
APELL1	VARCHAR(40)	YES
APELL2	VARCHAR(40)	YES
C_IDENTIDAD	VARCHAR(11)	NO
CHAPA_NUEVA	VARCHAR(50)	YES
CHAPA_VIEJA	VARCHAR(50)	YES
CLASE	VARCHAR(50)	YES
CLASE_DESC	VARCHAR(50)	YES
COLOR	VARCHAR(20)	YES
COLOR_DESC	VARCHAR(30)	YES
COLORP_DESC	VARCHAR(30)	YES
COMBUS_DESC	VARCHAR(50)	YES
COMBUSTIBLE	VARCHAR(50)	YES
DIRECCION	VARCHAR(70)	YES
ESTADO_CHAPA	VARCHAR(50)	YES
ESTADO_VEHICULO	VARCHAR(30)	YES
ID_REGVEH	INTEGER	NO
MARCA_VEHICULO	VARCHAR(15)	YES
MATRICULA	VARCHAR(20)	YES
MOT_VEHI_LIC	VARCHAR(50)	YES
NOMBRE	VARCHAR(40)	YES
NOMBRE2	VARCHAR(20)	YES
ORGA_EMPRESARIAL	VARCHAR(15)	YES
PAIS	VARCHAR(50)	YES
PROVINCIA	VARCHAR(40)	YES
PROVINCIA_PERSONA	VARCHAR(40)	YES
SECTOR	VARCHAR(15)	YES
SECTOR_DESC	VARCHAR(30)	YES
SEXO	VARCHAR(50)	YES
STVEH_DESC	VARCHAR(50)	YES
VIA_ADQ_DESC	VARCHAR(50)	YES

T_SAJO

Campo	Tipo	Nulo
ALIAS	VARCHAR(20)	YES
ALTURA	NUMBER	YES
C_IDENTIDAD	VARCHAR(11)	NO
CAUSA	VARCHAR(30)	YES
CLASIFICACION_NP	VARCHAR(50)	YES
CLASIFICACION_P	VARCHAR(30)	YES
COLOR_OJOS	VARCHAR(15)	YES
COLOR_PELO	VARCHAR(15)	YES
COMPOSTURA	VARCHAR(150)	YES
CONDUCTA	VARCHAR(50)	YES
DELITO	VARCHAR(30)	YES
ESTADO CIVIL	VARCHAR(15)	YES
FECHA_HECHO	DATE	YES
GRAVEDAD_DELITO	VARCHAR(50)	YES
ID_SAJO	VARCHAR(50)	NO
LUGAR_DELITO	VARCHAR(60)	YES
MEDIDA	VARCHAR(200)	YES
MEDIO	VARCHAR(50)	YES
MODO_OPERAR	VARCHAR(250)	YES
NIVEL_ESCOLARIDAD	VARCHAR(30)	YES
OBJETO	VARCHAR(60)	YES
OCUPACION	VARCHAR(40)	YES

T_SISDED

Campo	Tipo	Nulo
ADVERTENCIA	VARCHAR(500)	YES
C_IDENTIDAD	VARCHAR(11)	NO
CONDUCIDOS	VARCHAR(50)	YES
CONSEJO	VARCHAR(25)	YES
CONTRAVENCION	VARCHAR(40)	YES
DECICIONES	VARCHAR(50)	YES
DELITOS	VARCHAR(30)	NO
DENUNCIA	VARCHAR(600)	YES
DESCRIPCION	VARCHAR(300)	YES
DESCRIPCION_ADVERTENCIA	VARCHAR(400)	YES
DESCRIPCION_CONTRAVENCION	VARCHAR(500)	YES
DIRECCION	VARCHAR(50)	YES
EXPEDIENTE	VARCHAR(300)	YES
FECHA	VARCHAR(50)	YES
FECHA_DELITO	DATE	YES
FECHA_EDENTIFICACION	VARCHAR(50)	YES
FECHA_HECHO	DATE	YES
FECHA_MOTIVO	VARCHAR(50)	YES
GRAVEDAD_HECHO	VARCHAR(15)	YES
HECHO	VARCHAR(50)	YES
HORA_MOTIVO	VARCHAR(50)	YES
ID_SISDED	VARCHAR(50)	NO
ILEGALIDAD	VARCHAR(100)	YES
LUGAR_IDENTIFICACION	VARCHAR(50)	YES
MEDIDA	VARCHAR(300)	YES
MOTIVO	VARCHAR(40)	YES
MUNICIPIO	VARCHAR(20)	YES
REGISTROS	INTEGER	YES
RESPUESTA	VARCHAR(500)	YES
RESULTADO	VARCHAR(500)	YES
SINTESIS	VARCHAR(150)	YES
TIPO_DENUNCIA	VARCHAR(50)	YES

2.2.2 Análisis de dependencias funcionales

Perfilado de dependencia

Para realizar el análisis de las dependencias dentro de las tablas de origen de las fuentes de datos se realiza un estudio de las dependencias funcionales, en especial las transitivas, que existen en las mismas.

Para cada tabla se escribe en negrita el concepto que abstrae la dependencia y los campos que relaciona. El campo subrayado se refiere al campo del cual dependen los otros.

T_BPOLICIA

PERSONA (C_IDENTIDAD, COLOR_OJO, ESTATURA, DESCRIPCON_MUNICIPIO, PELO, PIEL)

POLICÍA (ID_BPOLICIA, IDENTIDAD, MUNICIPIO, UNIDAD)

ADVERTENCIA (ADVERTENCIA, CONTRAVENCION, DESCRIPCION_ADVERTENCIA)

T_ELECT

PERSONA (C_IDENTIDAD, APELL1, APELL2, DIRECCION, FECHA_NACIMIENTO, NOMBRE, SEXO)

CONSEJOS (CONSEJOS, NOMBRE_CONSEJOS)

MUNICIPIO (MUNICIPIO, NOMBRE_MUNICIPIO)

PROVINCIA (PROVINCIA, NOMBRE_PROVINCIA)

COLEGIOS (COLEGIOS, CDR, CIRCUNSCRIPCION, ZONA)

T_ARMAS DE FUEGO

PERSONA (C_IDENTIDAD, PROVINCIA)

MARCA (MARCA)

ARMA (ID_ARMAS DE FUEGO, ARMA, TIPO_ARMA, CALIBRE, CATEGORIA, DEST_ARMA_OCUPACIN, PROCEDENCIA_ARMA, SERIE, FECHA_ALTA_ARMA, FECHA_LICENCIA)

ESTADO_ ARMA (ID_ARMAS DE FUEGO , ESTADO1, ESTADO_ ARMA, MOTIVO_ESTADO, FECHA_ESTADO)

T_CIRCULADOS

PERSONA (C_IDENTIDAD, PROVINCIA, CIUDADANIA)

DENUNCIA (NUMERO_DENUNCIA, ANO_DENUNCIA, UNIDAD, DELITO, CATEGORIA)

CIRCULADOS (ID_CIRCULADOS, DIRECCION_CIRCULACION, FECHA_CIRCULACION, MUNICIPIO_CIRCULADO, PROVINCIA_CIRCULADA)

T_CUBAFIS

PERSONA (C_IDENTIDAD, APELL1, APELL2, MUNICIPIO, NOMBRE1, NOMBRE2, PROVINCIA, SEXO)

CUBAFIS (ID_CUBAFIS, FECHA_ACTUAL, FECHA_REGISTRO)

T_ESTANCIA_CASERO

HUESPED (PASAPORTE, APELL1_HUESPED, CIUDADANIA, FECHA_NACIMIENTO_HUSPED, NOMBRE_HUSPED)

CASERO (C_IDENTIDAD_CASERO, CASERO)

ESTANCIA (ID_ESTANCIA_CASERO, DIRCCION_ESTANCIA, FECHA_FIN, FECHA_INICIO, , MUNICIPIO_ESTANCIA)

T_EXPECOND

PERSONA (C_IDENTIDAD, APELL1, APELL2, FECHA_NACIMIENTO, NOMBRE, NOMBRE2, SEXO, PROVINCIA, MUNICIPIO)

LICENCIA (LICENCIA, ESTADO_LICENCIA, FECHA_VENC_LICENCIA DESCRIPCION_LIMITANTE, DESCRIPCION_TRAMO)

EXPECOND (ID_EXPECOND, FECHA_EXP, FECHA_VENC)

T_REGULADOS

PERSONA (C_IDENTIDAD, PROFESION

CATEGORIA (CATEGORIA, DESCRIPCION_CATEGORIA)

REGULADOS (ID_REGULADOS, FECHA_ALTA, REGULADOS_VITALES)

T_REGVEH

PERSONA (C_IDENTIDAD, APELL1, APELL2, DIRECCION, NOMBRE, NOMBRE2, PAIS, PROVINCIA_PERSONA, SEXO)

CHAPA (CHAPA_NUEVA, CHAPA_VIEJA, ESTADO_CHAPA)

CLASE (CLASE, CLASE_DESC)

COMBUSTIBLE (COMBUSTIBLE, COMBUS_DESC)

VEHICULO (MATRICULA, ESTADO_VEHICULO, COLOR, COLOR_DESC, COLORP_DESC, ID_REGVEH, MARCA_VEHICULO, MOT_VEHI_LIC, STVEH_DESC, VIA_ADQ_DESC)

ORGA_EMPRESARIAL (ORGA_EMPRESARIAL, PROVINCIA)

SECTOR (SECTOR, SECTOR_DESC)

T_SAJO

PERSONA (C_IDENTIDAD, ALIAS, ALTURA, COLOR_OJOS, COLOR_PELO, COMPOSTURA, ESTADO CIVIL, NIVEL_ESCOLARIDAD, OCUPACION, CONDUCTA, MEDIDA)

DELITO (ID_SAJO , DELITO, CAUSA, GRAVEDAD_DELITO, LUGAR_DELITO, MEDIO, MODO_OPERAR, OBJETO, FECHA_HECHO)

CLASIFICACION (ID_SAJO , CLASIFICACION_NP, CLASIFICACION_P)

T_SISDED

ADVERTENCIA (ADVERTENCIA, DESCRIPCION_ADVERTENCIA)

PERSONA (C_IDENTIDAD, DIRECCION, MUNICIPIO, CONSEJO)

CONDUCIDOS (C_IDENTIDAD, CONDUCIDOS)

CONTRAVENCION (CONTRAVENCION, DESCRIPCION_CONTRAVENCION)

DELITOS (DELITOS, FECHA_DELITO)

HECHO (HECHO, FECHA_HECHO, GRAVEDAD_HECHO)

EXPEDIENTE (EXPEDIENTE, FECHA, DESCRIPCION, LUGAR_IDENTIFICACION, DECICIONES, FECHA_EDENTIFICACION, ID_SISDED, ILEGALIDAD, MEDIDA, REGISTROS, RESPUESTA, RESULTADO, SINTESIS)

DENUNCIA (DENUNCIA, TIPO_DENUNCIA)

MOTIVO (MOTIVO, FECHA_MOTIVO, HORA_MOTIVO)

2.2.3 Mapeo de los datos

Para mapear los datos de las fuentes de datos hacia el área intermedia (staging área) se una elabora un mapa lógico de datos. Un mapa lógico de datos es una estructura, por lo general una tabla con las características de las fuentes de información, algunos elementos que podría contener esta tabla son:

- Nombre de la tabla destino
- Nombre de la columna destino
- Tipo de tabla
- SCD (Grado del cambio de dimensión)
- Base de datos fuente
- Nombre de la tabla origen
- Nombre de la columna destino
- Transformaciones

Para elaborar el mapa lógico de datos, se procede por diversas fases:

- La fase de exploración y descubrimiento de información, donde se colecta y documenta la información de los diversos sistemas fuentes de datos, además de ello se comienza con el seguimiento de los mismos.
- La fase de análisis del contenido de datos, donde se navega a través de la fuentes de datos y se recaba información acerca de valorares nulos, codificación y estructura de los esquemas de datos.
- La fase de recolección de reglas de negocio
- La fase de integración de fuentes de datos heterogéneas

Con los mapas lógicos obtenidos se facilita el proceso de extracción de los datos hacia el área intermedia ya que se puede conocer cuáles son los datos que se tomaran y de donde extraerlos. Permitirá tomar decisiones sobre datos similares desde fuentes heterogéneas y determinar las posibles transformaciones en etapas posteriores.

Como resultado de la aplicación de esta técnica se obtuvieron los siguientes mapas de datos.

Tabla origen: T_BPOLICIA

Fuente			Destino			
Campo	**Tipo**	**Nulo**	**Campo**	**Tabla**	**Tipo**	**Nulo**
ADVERTENCIA	VARCHAR(50)	YES	ADVERTENCIA	ESTACION_POLICIA_PERSONA	CHAR(1000)	NO
C_IDENTIDAD	VARCHAR(11)	NO	CARNET_IDENTIDAD	PERSONA	CHAR(11)	NO
COLOR_OJO	VARCHAR(20)	YES	COLOR	COLOR_OJOS	CHAR(20)	NO
CONTRAVENCION	VARCHAR(50)	YES	CONTRAVENCION	CONTRAVENCION	CHAR(1000)	NO
DESCRIPCION_ADVERTENCIA	VARCHAR(500)	YES	ADVERTENCIA	ESTACION_POLICIA_PERSONA	CHAR(1000)	NO
DESCRIPCON_MUNICIPIO	VARCHAR(20)	YES	NOMBRE	MUNICIPIO	CHAR(20)	NO
ESTATURA	FLOAT	YES	ESTATURA	PERSONA	FLOAT	NO
ID_BPOLICIA	INTEGER	NO				NO
IDENTIDAD	VARCHAR(30)	YES	IDENTIDA	POLICIA	CHAR(10)	NO
MUNICIPIO	VARCHAR(50)	YES	NOMBRE	MUNICIPIO	CHAR(50)	NO
PELO	VARCHAR(20)	YES	PELO	PELO	CHAR(20)	NO
PIEL	VARCHAR(20)	YES	PIEL	PIEL	CHAR(20)	NO
UNIDAD	INTEGER	YES	NUMERO	ESTACION_POLICIA	NUMBER	NO

Tabla origen: T_ELECT

Fuente			Destino			
Campo	**Tipo**	**Nulo**	**Campo**	**Tabla**	**Tipo**	**Nulo**
APELL1	VARCHAR(30)	YES	APELLIDOS	PERSONA	CHAR(50)	NO
APELL2	VARCHAR(30)	YES	APELLIDOS	PERSONA	CHAR(50)	NO
C_IDENTIDAD	VARCHAR(11)	NO	CARNET_IDENTIDAD	PERSONA	CHAR(11)	NO
CDR	VARCHAR(50)	YES				NO
CIRCUNSCRIPCION	INTEGER	YES	CIRCUNSCRIPCION	RESIDENCIA	NUMBER	NO
COLEGIOS	VARCHAR(60)	YES				NO
CONSEJOS	VARCHAR(20)	YES	NOMBRE	CONSEJO_POPULAR	CHAR(30)	NO
DIRECCION	VARCHAR(50)	YES	DIRECCION	RESIDENCIA	CHAR(60)	NO
FECHA_NACIMIENTO	DATE	YES	FECHA_NAC	PERSONA	DATE	NO
MUNICIPIO	VARCHAR(15)	YES	NOMBRE	MUNICIPIO	CHAR(20)	NO
NOMBRE	VARCHAR(20)	YES	NOMBRE	PERSONA	CHAR(30)	NO
NOMBRE_CONSEJOS	VARCHAR(20)	YES	NOMBRE	CONSEJO_POPULAR	CHAR(30)	NO
NOMBRE_MUNICIPIO	VARCHAR(15)	YES	NOMBRE	MUNICIPIO	CHAR(30)	NO
NOMBRE_PROVINCIA	VARCHAR(20)	YES	NOMBRE	PROVINCIA	CHAR(10)	NO
PROVINCIA	VARCHAR(20)	YES	NOMBRE	PROVINCIA	CHAR(10)	NO
SEXO	VARCHAR(10)	YES	SEXO	PERSONA	CHAR(1)	NO
ZONA	VARCHAR(50)	YES	ZONA	RESIDENCIA	NUMBER	NO

Tabla origen: T_ARMAS DE FUEGO

F u e n t e			**D e s t i n o**			
Campo	**Tipo**	**Nulo**	**Campo**	**Tabla**	**Tipo**	**Nulo**
C_IDENTIDAD	VARCHAR(11)	NO	CARNET_IDENTIDAD	PERSONA	CHAR(11)	NO
CALIBRE	INTEGER	YES	CALIBRE	ARMA_FUEGO	NUMBER	NO
CATEGORIA	VARCHAR(30)	YES	CATEGORIA	CATEGORIA_ARMA	CHAR(10)	NO
DEST_ARMA_OCUPACIN	VARCHAR(30)	YES	DESTINO	DESTINO_ARMA	CHAR(10)	NO
ESTADO_ ARMA	VARCHAR(40)	YES	ESTADO	ESTADO_ARMA	CHAR(20)	NO
ESTADO1	VARCHAR(40)	YES	ESTADO	ESTADO_ARMA	CHAR(20)	NO
FECHA_ALTA_ARMA	DATE	YES	FECHA_ALTA	ARMA_FUEGO	DATE	NO
FECHA_ESTADO	DATE	YES	FECHA	ESTADO_ARMA	DATE	NO
FECHA_LICENCIA	DATE	YES	FECHA LICENCIA	ARMA_FUEGO	DATE	NO
ID_ARMAS DE FUEGO	INTEGER	NO	ID_ARMA	ARMA_FUEGO	NUMBER	NO
MARCA	VARCHAR(30)	YES	MARCA	ARMA_FUEGO	CHAR(20)	NO
MOTIVO_ESTADO	VARCHAR(50)	YES	MOTIVO	ESTADO_ARMA	CHAR(100)	NO
PROCEDENCIA_ARMA	VARCHAR(40)	YES	PROCEDENCIA	PROCEDENCIA_ARMA	CHAR(30)	NO
PROVINCIA	VARCHAR(50)	YES	NOMBRE	PROVINCIA	CHAR(10)	NO
SERIE	INTEGER	YES	SERIE	ARMA_FUEGO	CHAR(20)	NO
TIPO_ARMA	VARCHAR(50)	YES	TIPO	TIPO_ARMA	CHAR(20)	NO

Tabla origen: T_CIRCULADOS

Fuente			Destino			
Campo	**Tipo**	**Nulo**	**Campo**	**Tabla**	**Tipo**	**Nulo**
ANO_DENUNCIA	VARCHAR(50)	YES	FECHA	DENUNCIA-PERSONA	DATE	NO
C_IDENTIDAD	VARCHAR(11)	NO	CARNET_IDENTIDAD	PERSONA	CHAR(11)	NO
CATEGORIA	VARCHAR(50)	YES	CATEGORIA	CATEGORIA_DELITO	CHAR(20)	NO
CIUDADANIA	VARCHAR(50)	YES	CIUDADANIA	CIUDADANIA	CHAR(20)	NO
DELITO	VARCHAR(30)	YES	DELITO	DELITO	CHAR(20)	NO
DIRECCION_CIRCULACION	VARCHAR(50)	YES	DIRECCION	RESIDENCIA		NO
FECHA_CIRCULACION	DATE	YES	FECHA	CIRCULADO	DATE	NO
ID_CIRCULADOS	INTEGER	NO				NO
MUNICIPIO_CIRCULADO	VARCHAR(40)	YES	NOMBRE	MUNICIPIO	CHAR(30)	NO
NUMERO_DENUNCIA	INTEGER	YES	NUMERO	CIRCULADO	NUMBER	NO
PROVINCIA	VARCHAR(30)	YES	NOMBRE	PROVINCIA	CHAR(10)	NO
PROVINCIA_CIRCULADA	VARCHAR(50)	YES	NOMBRE	PROVINCIA	CHAR(10)	NO
UNIDAD	VARCHAR(50)	YES	NUMERO	ESTACION_POLICIA	NUMBER	NO

Tabla origen: T_CUBAFIS

Fuente			Destino			
Campo	**Tipo**	**Nulo**	**Campo**	**Tabla**	**Tipo**	**Nulo**
APELL1	VARCHAR(40)	YES	APELLIDOS	PERSONA	CHAR(50)	NO
APELL2	VARCHAR(40)	YES	APELLIDOS	PERSONA	CHAR(50)	NO
C_IDENTIDAD	VARCHAR(11)	NO	CARNET_IDENTIDAD	PERSONA	CHAR(11)	NO
FECHA_ACTUAL	DATE	YES	FECHA	REGISTROS	DATE	NO
FECHA_REGISTRO	DATE	YES	FECHA	REGISTROS_PERSONA	DATE	NO
ID_CUBAFIS	VARCHAR(30)	NO				NO
MUNICIPIO	VARCHAR(50)	YES	NOMBRE	MUNICIPIO	CHAR(20)	NO
NOMBRE1	VARCHAR(20)	YES	NOMBRE	PERSONA	CHAR(30)	NO
NOMBRE2	VARCHAR(20)	YES	NOMBRE	PERSONA	CHAR(30)	NO
PROVINCIA	VARCHAR(50)	YES	NOMBRE	PROVINCIA	CHAR(10)	NO
SEXO	VARCHAR(50)	YES	SEXO	PERSONA	CHAR(1)	NO

Tabla origen: T_ESTANCIA_CASERO

Fuente			Destino			
Campo	**Tipo**	**Nulo**	**Campo**	**Tabla**	**Tipo**	**Nulo**
APELL1_HUESPED	VARCHAR(40)	YES	APELLIDOS	PERSONA	CHAR(50)	NO
C_IDENTIDAD	VARCHAR(11)	NO	CARNET_IDENTIDAD	PERSONA	CHAR(11)	NO
C_IDENTIDAD_CASERO	VARCHAR(11)	YES	CARNET_IDENTIDAD	PERSONA	CHAR(11)	NO
CIUDADANIA	VARCHAR(50)	YES	CIUDADANIA	CIUDADANIA	CHAR(20)	NO
DIRCCION_ESTANCIA	VARCHAR(50)	YES	DIRECCION	RESIDENCIA	CHAR(60)	NO
FECHA_FIN	DATE	YES	FECHA	HOSPEDAJE_EXTRANJERO	DATE	NO
FECHA_INICIO	DATE	YES	FECHA	HOSPEDAJE_EXTRANJERO	DATE	NO
FECHA_NACIMIENTO_HUSPED	DATE	YES	FECHA_NAC	PERSONA	DATE	NO
ID_ESTANCIA_CASERO	VARCHAR(50)	NO				NO
MUNICIPIO_ESTANCIA	VARCHAR(40)	YES	NOMBRE	MUNICIPIO	CHAR(20)	NO
NOMBRE_HUSPED	VARCHAR(50)	YES	ID_EXTRANJERO	EXTRANJERO	CHAR(10)	NO
PASAPORTE	VARCHAR(50)	YES	PASAPORTE	EXTRANJERO	CHAR(20)	NO

Tabla de origen: T_EXPECOND

Fuente			Destino			
Campo	**Tipo**	**Nulo**	**Campo**	**Tabla**	**Tipo**	**Nulo**
APELL1	VARCHAR(40)	YES	APELLIDOS	PERSONA	CHAR(30)	NO
APELL2	VARCHAR(40)	YES	APELLIDOS	PERSONA	CHAR(30)	NO
C_IDENTIDAD	VARCHAR(11)	NO	CARNET_IDENTIDAD	PERSONA	CHAR(11)	NO
CHAPA	VARCHAR(15)	YES	MATRICULA	MATRICULA	CHAR(10)	NO
DESCRIPCION_LIMITANTE	VARCHAR(200)	YES	LIMITANTE	EXPEDIENTE_CONDUCCION	CHAR(100)	NO
DESCRIPCION_TRAMO	VARCHAR(150)	YES	TRAMO	EXPEDIENTE_CONDUCCION	CHAR(100)	NO
DESCRIPCION_UNIDAD	INTEGER	YES	NUMERO	ESTACION_POLICIA	NUMBER	NO
ESTADO_LICENCIA	VARCHAR(50)	YES	ESTADO_LICENCIA	EXPEDIENTE_CONDUCCION	CHAR(10)	NO
FECHA_EXP	DATE	YES	FECHA_EXPEDIDA	EXPEDIENTE_CONDUCCION	DATE	NO
FECHA_VENC	DATE	YES	FECHA_VENCIMIENTO	EXPEDIENTE_CONDUCCION	DATE	NO
FECHA_VENC_LICENCIA	DATE	YES	FECHA_VENCIMIENTO	EXPEDIENTE_CONDUCCION	DATE	NO
ID_EXPECOND	INTEGER	NO	ID_EXPEDIENTE	EXPEDIENTE_CONDUCCION	CHAR(10)	NO
LICENCIA	VARCHAR(20)	YES	LICENCIA	EXPEDIENTE_CONDUCCION	CHAR(10)	NO
MUNICIPIO	VARCHAR(30)	YES	NOMBRE	MUNICIPIO	CHAR(20)	NO
NOMBRE	VARCHAR(30)	YES	NOMBRE	PERSONA	CHAR(30)	NO
NOMBRE2	VARCHAR(20)	YES	NOMBRE	PERSONA	CHAR(30)	NO
PROVINCIA	VARCHAR(20)	YES	NOMBRE	PROVINCIA	CHAR(10)	NO

Tabla origen: T_REGULADOS

Fuente			Destino			
Campo	**Tipo**	**Nulo**	**Campo**	**Tabla**	**Tipo**	**Nulo**
C_IDENTIDAD	VARCHAR(11)	YES	CARNET_IDENTIDAD	PERSONA	CHAR(11)	NO
CATEGORIA	VARCHAR(50)	YES	CATEGORIA	CATEGORIA_REGULADO	CHAR(20)	NO
DESCRIPCION_CATEGORIA	VARCHAR(200)	YES	DESCRIPCION	CATEGORIA_REGULADO	CHAR(1000)	NO
FECHA_ALTA	DATE	YES	FECHA_ALTA	REGULADO	DATE	NO
ID_REGULADOS	VARCHAR(20)	YES	CARNET_IDENTIDAD	PERSONA	CHAR(11)	NO
PROFESION	VARCHAR(30)	YES	PROFESION	PROFESION	CHAR(20)	NO
REGULADOS_VITALES	VARCHAR(10)	YES	VITALICIO	REGULADO	CHAR(1)	NO

Tabla origen: T_REGVEH

Fuente			Destino			
Campo	**Tipo**	**Nulo**	**Campo**	**Tabla**	**Tipo**	**Nulo**
APELL1	VARCHAR(40)	YES	APELLIDOS	PERSONA	CHAR(30)	NO
APELL2	VARCHAR(40)	YES	APELLIDOS	PERSONA	CHAR(30)	NO
C_IDENTIDAD	VARCHAR(11)	NO	CARNET_IDENTIDAD	PERSONA	CHAR(11)	NO
CHAPA_NUEVA	VARCHAR(50)	YES	MATRICULA	MATRICULA	CHAR(10)	NO
CHAPA_VIEJA	VARCHAR(50)	YES	MATRICULA	MATRICULA	CHAR(10)	NO
CLASE	VARCHAR(50)	YES	CLASE	CLASE_VEHICULO	CHAR(10)	NO
CLASE_DESC	VARCHAR(50)	YES	DESCRIPCION	CLASE_VEHICULO	CHAR(500)	NO
COLOR	VARCHAR(20)	YES	COLOR	VEHICULO	CHAR(10)	NO
COLOR_DESC	VARCHAR(30)	YES	COLOR	VEHICULO	CHAR(10)	NO
COLORP_DESC	VARCHAR(30)	YES	COLOR	VEHICULO	CHAR(10)	NO
COMBUS_DESC	VARCHAR(50)	YES	DESCRIPCION	COMBUSTIBLE	CHAR(500)	NO
COMBUSTIBLE	VARCHAR(50)	YES	COMBUSTIBLE	COMBUSTIBLE	CHAR(10)	NO
DIRECCION	VARCHAR(70)	YES	DIRECCION	RESIDENCIA	CHAR(60)	NO
ESTADO_CHAPA	VARCHAR(50)	YES	ESTADO_MATRICULA	VEHICULO_MATRICULA	CHAR(50)	NO
ESTADO_VEHICULO	VARCHAR(30)	YES	ESTADO	ESTADO_VEHICULO	CHAR(20)	NO
ID_REGVEH	INTEGER	NO	ID_VEHICULO	VEHICULO	NUMBER	NO
MARCA_VEHICULO	VARCHAR(15)	YES	MARCA	MARCA_VEHICULO	CHAR(20)	NO
MATRICULA	VARCHAR(20)	YES	MATRICULA	MATRICULA	CHAR(10)	NO
MOT_VEHI_LIC	VARCHAR(50)	YES	LICENCIA_MOTOR	VEHICULO	CHAR(10)	NO
NOMBRE	VARCHAR(40)	YES	NOMBRE	PERSONA	CHAR(30)	NO
NOMBRE2	VARCHAR(20)	YES	NOMBRE	PERSONA	CHAR(30)	NO
ORGA_EMPRESARIAL	VARCHAR(15)	YES	SECTOR	SECTOR	CHAR(20)	NO

PAIS	VARCHAR(50)	YES	PAIS	PAIS_PROCEDENCIA	CHAR(20)	NO
PROVINCIA	VARCHAR(40)	YES	NOMBRE	PROVINCIA	CHAR(10)	NO
PROVINCIA_PERSONA	VARCHAR(40)	YES	NOMBRE	PROVINCIA	CHAR(10)	NO
SECTOR	VARCHAR(15)	YES	SECTOR	SECTOR	CHAR(20)	NO
SECTOR_DESC	VARCHAR(30)	YES	SECTOR	SECTOR	CHAR(20)	NO
SEXO	VARCHAR(50)	YES	SEXO	PERSONA	CHAR(1)	NO
STVEH_DESC	VARCHAR(50)	YES	ESTADO	VEHICULO	CHAR(500)	NO
VIA_ADQ_DESC	VARCHAR(50)	YES	VIA_ADQUISICION	VEHICULO	CHAR(20)	NO

Tabla origen: T_SAJO

Fuente			Destino			
Campo	**Tipo**	**Nulo**	**Campo**	**Tabla**	**Tipo**	**Nulo**
ALIAS	VARCHAR(20)	YES	ALIAS	PERSONA	CHAR(200)	NO
ALTURA	NUMBER	YES	ESTATURA	PERSONA	FLOAT	NO
C_IDENTIDAD	VARCHAR(11)	NO	CARNET_IDENTIDAD	PERSONA	CHAR(11)	NO
CAUSA	VARCHAR(30)	YES	CAUSA	DELITO	CHAR(20)	NO
CLASIFICACION_NP	VARCHAR(50)	YES	CLASIFICACION	CLASIFICACION_NP	CHAR(20)	NO
CLASIFICACION_P	VARCHAR(30)	YES	CLASIFICACION_P	CLASIFICACION_P	CHAR(20)	NO
COLOR_OJOS	VARCHAR(15)	YES	COLOR	COLOR_OJOS	CHAR(20)	NO
COLOR_PELO	VARCHAR(15)	YES	PELO	PELO	CHAR(20)	NO
COMPOSTURA	VARCHAR(150)	YES	COMPOSTURA	PERSONA	CHAR(200)	NO
CONDUCTA	VARCHAR(50)	YES				NO
DELITO	VARCHAR(30)	YES	DELITO	DELITO	CHAR(20)	NO
ESTADO CIVIL	VARCHAR(15)	YES	ESTADO	ESTADO_CIVIL	CHAR(20)	NO
FECHA_HECHO	DATE	YES	FECHA	HECHO	DATE	NO
GRAVEDAD_DELITO	VARCHAR(50)	YES	GRAVEDAD	DELITO	CHAR(20)	NO
ID_SAJO	VARCHAR(50)	NO				NO
LUGAR_DELITO	VARCHAR(60)	YES	LUGAR	DELITO	CHAR(50)	NO
MEDIDA	VARCHAR(200)	YES	MEDIDA	DELITO	CHAR(500)	NO
MEDIO	VARCHAR(50)	YES	MEDIO	DELITO	CHAR(20)	NO
MODO_OPERAR	VARCHAR(250)	YES	MODO_OPERAR	DELITO	CHAR(500)	NO
NIVEL_ESCOLARIDAD	VARCHAR(30)	YES	ESCOLARIDAD	ESCOLARIDAD	CHAR(10)	NO
OBJETO	VARCHAR(60)	YES	OBJETO	DELITO	CHAR(100)	NO

OCUPACION	VARCHAR(40)	YES	PROFESION	PROFESION	CHAR(20)	NO

Tabla de origen: T_SISDED

Fuente			Destino			
Campo	**Tipo**	**Nulo**	**Campo**	**Tabla**	**Tipo**	**Nulo**
ADVERTENCIA	VARCHAR(500)	YES	ADVERTENCIA	ESTACION_POLICIA_PERSONA	CHAR(1000)	NO
C_IDENTIDAD	VARCHAR(11)	NO	CARNET_IDENTIDAD	PERSONA	CHAR(11)	NO
CONDUCIDOS	VARCHAR(50)	YES	CARNET_IDENTIDAD	ESTACION_POLICIA_PERSONA	CHAR(11)	NO
CONSEJO	VARCHAR(25)	YES	NOMBRE	CONSEJO_POPULAR	CHAR(30)	NO
CONTRAVENCION	VARCHAR(40)	YES	CONTRAVENCION	CONTRAVENCION	CHAR(1000)	NO
DECICIONES	VARCHAR(50)	YES	DECISION	DENUNCIA_PERSONA	CHAR(500)	NO
DELITOS	VARCHAR(30)	NO	DELITO	DELITO	CHAR(20)	NO
DENUNCIA	VARCHAR(600)	YES	DENUNCIA	DENUNCIA	CHAR(1000)	NO
DESCRIPCION	VARCHAR(300)	YES	DESCRIPCION	EXPEDIENTE	CHAR(500)	NO
DESCRIPCION_ADVERTENCIA	VARCHAR(400)	YES	ADVERTENCIA	ESTACION_POLICIA_PERSONA	CHAR(1000)	NO
DESCRIPCION_CONTRAVENCION	VARCHAR(500)	YES	CONTRAVENCION	CONTRAVENCION	CHAR(1000)	NO
DIRECCION	VARCHAR(50)	YES	DIRECCION	RESIDENCIA	CHAR(60)	NO

EXPEDIENTE	VARCHAR(300)	YES	DESCRIPCION	EXPEDIENTE	CHAR(500)	NO
FECHA	VARCHAR(50)	YES	FECHA	EXPEDIENTE	DATE	NO
FECHA_DELITO	DATE	YES	FECHA	DELITO_PERSONA	DATE	NO
FECHA_EDENTIFICACION	VARCHAR(50)	YES	FECHA	EXPEDIENTE	DATETIME	NO
FECHA_HECHO	DATE	YES	FECHA	HECHO	DATE	NO
FECHA_MOTIVO	VARCHAR(50)	YES	FECHA_MOTIVO	MEDIDA_PERSONA	DATETIME	NO
GRAVEDAD_HECHO	VARCHAR(15)	YES	GRAVEDAD	HECHO	CHAR(10)	NO
HECHO	VARCHAR(50)	YES	HECHO	HECHO	CHAR(50)	NO
HORA_MOTIVO	VARCHAR(50)	YES	FECHA	HECHO	DATETIME	NO
ID_SISDED	VARCHAR(50)	NO				NO
ILEGALIDAD	VARCHAR(100)	YES	ILEGALIDAD	ILEGALIDAD	CHAR(200)	NO
LUGAR_IDENTIFICACION	VARCHAR(50)	YES	LUGAR	DELITO	CHAR(50)	NO
MEDIDA	VARCHAR(300)	YES	MEDIDA	DELITO	CHAR(500)	NO
MOTIVO	VARCHAR(40)	YES	MOTIVO	MEDIDA	CHAR(50)	NO
MUNICIPIO	VARCHAR(20)	YES	NOMBRE	MUNICIPIO	CHAR(30)	NO
REGISTROS	INTEGER	YES	ID_REGISTRO	REGISTRO	NUMBER	NO
RESPUESTA	VARCHAR(500)	YES	DESICIONES	REGISTRO	CHAR(1000)	NO
RESULTADO	VARCHAR(500)	YES	RESULTADO	REGISTRO	CHAR(1000)	NO
SINTESIS	VARCHAR(150)	YES	SINTESIS	EXPEDIENTE	CHAR(200)	NO
TIPO_DENUNCIA	VARCHAR(50)	YES	TIPO_DENUNCIA	TIPO_DENUNCIA	CHAR(20)	NO

2.3 El área intermedia (staging área)

Con el objeto de minimizar al máximo nivel los posibles errores o problemas en la fase de carga de los procesos ETL, normalmente se reserva un área de disco para poder recuperar los datos por etapas. Por lo que se puede afirmar que esta área intermedia (staging) está estrechamente relacionado tanto con el gerenciamiento como con la recuperación de datos.

Un área de stage (se puede traducir como área de prueba o área de ensayo), es un área intermedia de almacenamiento de datos utilizada para el procesamiento de los mismos durante procesos de extracción, transformación y carga (ETL). Esta área se encuentra entre la fuente de los datos y su destino, que a menudo son almacenes de datos, datamarts u otros repositorios de datos.

Las áreas de stage de datos son a menudo de naturaleza transitoria, su contenido se borrará antes de ejecutar un proceso de ETL o inmediatamente después de haberlo finalizado con éxito. Aunque existen arquitecturas de área stage diseñadas para mantener los datos durante largos períodos de tiempo con la finalidad de mantener un archivo de los mismos o para poder resolver problemas detectados a posteriori.

Las áreas de stage se pueden implementar en forma de tablas de bases de datos relacionales, archivos de texto plano (como archivos XML o CSV) o archivos binarios propietarios almacenados en un determinado sistema de archivos. Las arquitecturas para área de stage varían en complejidad, desde un conjunto de simples tablas relacionales en una base de datos de destino hasta instancias de bases de datos auto-contenidas o sistemas de archivos. A pesar de que los sistemas de origen y de destino de un proceso ETL son a menudo bases de datos relacionales, no es necesario las zonas de stage que se ubican entre ambos también lo sean.

Funcionamiento del staging se basa en dos pasos. En primer lugar, los datos son volcados por bloques o etapas y de forma independiente en un área del disco denominada staging área. Posteriormente, se cargan los datos desde la staging área a su lugar o sistema de destino (datamart).

Ventajas de utilizar una staging área:

- Permite independizar el proceso de carga por bloques o etapas. Lo cual es muy útil y práctico cuando se trabaja con muchos datos, ya que evita tener que reiniciar el proceso entero en caso de error o avería. Por ejemplo, si se produjese un corte eléctrico, solo habría que repetir el volcado de datos del bloque específico en el que se ha producido la incidencia, estando el resto de información a buen recaudo y segura en el área de staging.
- Si se implementa correctamente, posibilita reiniciar las distintas fases del proceso ETL de manera independiente. Esto significa que si, por ejemplo, falla el proceso de transformación, bastaría con volver a repetir esta fase, pero no sería necesario repetir la etapa anterior: la de extracción.
- La compilación de los distintos bloques o etapas del proceso de staging puede incluso adaptarse a las necesidades de los clientes, aunque siempre que esté contemplado previamente en el proceso general del ETL.
- Al tratarse de un disco físicamente independiente, en ningún caso afecta ni ralentiza otros procesos del sistema.

Las zonas stage pueden proporcionar beneficios diversos, pero la principal motivación para su uso es aumentar la eficiencia de los procesos ETL, garantizar la integridad de los datos y apoyar ciertas operaciones que aseguren la calidad de los mismos. Las funciones de un área de stage son las siguientes:

1. Consolidación de datos: Una de las principales funciones de un área de stage es la consolidación de datos de múltiples sistemas de origen. Para ello el área de stage actúa como un gran "cubo" en el que los datos de varios sistemas de origen se ubican temporalmente para su posterior procesamiento. Adicionalmente, los datos del área de stage se suelen caracterizar con ciertos metadatos para identificar la fuente de origen, el momento (fecha/hora) en que los datos fueron cargados en esta zona u otra información que se considere relevante.
2. Alineación: La Alineación de datos consiste en la estandarización de estos a través de los múltiples sistemas de origen y la validación de las

relaciones entre los registros y elementos de datos de diferentes fuentes. Esta función está estrechamente relacionada con la administración de datos maestros,5 ya que da soporte a este tipo de gestiones.

3. Minimizar la contención: Tanto el área de stage como los procesos ETL que apoya, a menudo se diseñan con el objetivo de minimizar la "discordia" en los sistemas de origen. A veces resulta más eficiente copiar los datos requeridos de un sistema de origen a la zona de stage de un golpe que tratar de recuperar únicamente registros individuales o pequeños conjuntos de registros.
4. El primer método, el área stage, se aprovecha de eficiencia técnica de las tecnologías de transmisión de datos, la reducción de los gastos generales a través de minimizar la necesidad de romper y volver a establecer las conexiones con los sistemas de origen y la optimización de la gestión de bloqueo de concurrencia en los sistemas de origen multi-usuario. Por su parte, los procesos ETL ejercen un alto grado de control sobre los problemas de concurrencia durante el procesamiento.
5. Planificación independiente de objetivos múltiples: El alojamiento de datos en un área de stage permite planificar de forma independiente, en cualquier momento, las operaciones de procesamiento de los mismos, pudiéndose realizar dichas operaciones cuando los diversos objetivos del negocio lo requieran. En algunos casos, los datos se podrían llevar a la zona de stage en diferentes momentos, para luego procesarlos todos a la vez. Esta situación puede ocurrir, por ejemplo, cuando el normal desempeño de la empresa se realiza a través de múltiples zonas horarias. En otros casos, los datos se pueden cargar en el área de stage para ser procesados en diferentes momentos. El área de stage también se puede utilizar para enviar datos a múltiples sistemas de destino en diferentes momentos; por ejemplo, los datos operacionales diarios podrían ser enviados a un almacén operacional de datos (ODS), mientras que los mismos datos se podrían enviar mensualmente de forma agregada a un almacén de datos.
6. Detección de cambios: El área de stage permite realizar una detección de cambios eficaz frente a los sistemas de destino. Esta funcionalidad es

particularmente útil cuando los sistemas de origen no soportan formas fiables de detección de cambios, tales como el sellado de tiempo (timestamping) impuesto por el sistema, el control de cambios (trazabilidad) o captura de datos modificados (CDC, change data capture).

7. Limpieza de datos: La limpieza de datos consiste en la identificación y eliminación (o actualización) de datos no válidos de los sistemas de origen. El proceso ETL, utilizando el área de stage, se puede utilizar para implementar la lógica de negocio que permita identificar y manejar los datos "no válidos". Los datos no válidos se identifican a menudo mediante una combinación de reglas de negocio y ciertas limitaciones técnicas, las cuales, se pueden integrar en la estructura del área de stage (como por ejemplo, restricciones de tabla en una base de datos relacional) para hacer cumplir las reglas de validez de los datos.
8. Cálculo de agregados: El pre-cálculo de valores agregados, otros tipos de cálculos y la aplicación de una lógica de negocio compleja puede hacerse en un área de stage para dar soporte a acuerdos de nivel de servicio altamente sensibles (o SLA, service-level agreement) o para la presentación de informes de resumen en los sistemas de destino.
9. Archivo de datos y resolución de problemas: Un área de stage da soporte y permite realizar el archivo de datos. En este escenario esta zona se puede utilizar para mantener los registros históricos durante el proceso de carga, o se puede utilizar para enviar datos a una estructura de archivos de destino. Además los datos pueden conservarse durante largos periodos de tiempo para permitir resolver los problemas técnicos que puedan surgir en las operaciones ETL.

2.3.1 Staging área de las fuentes de datos del MININT

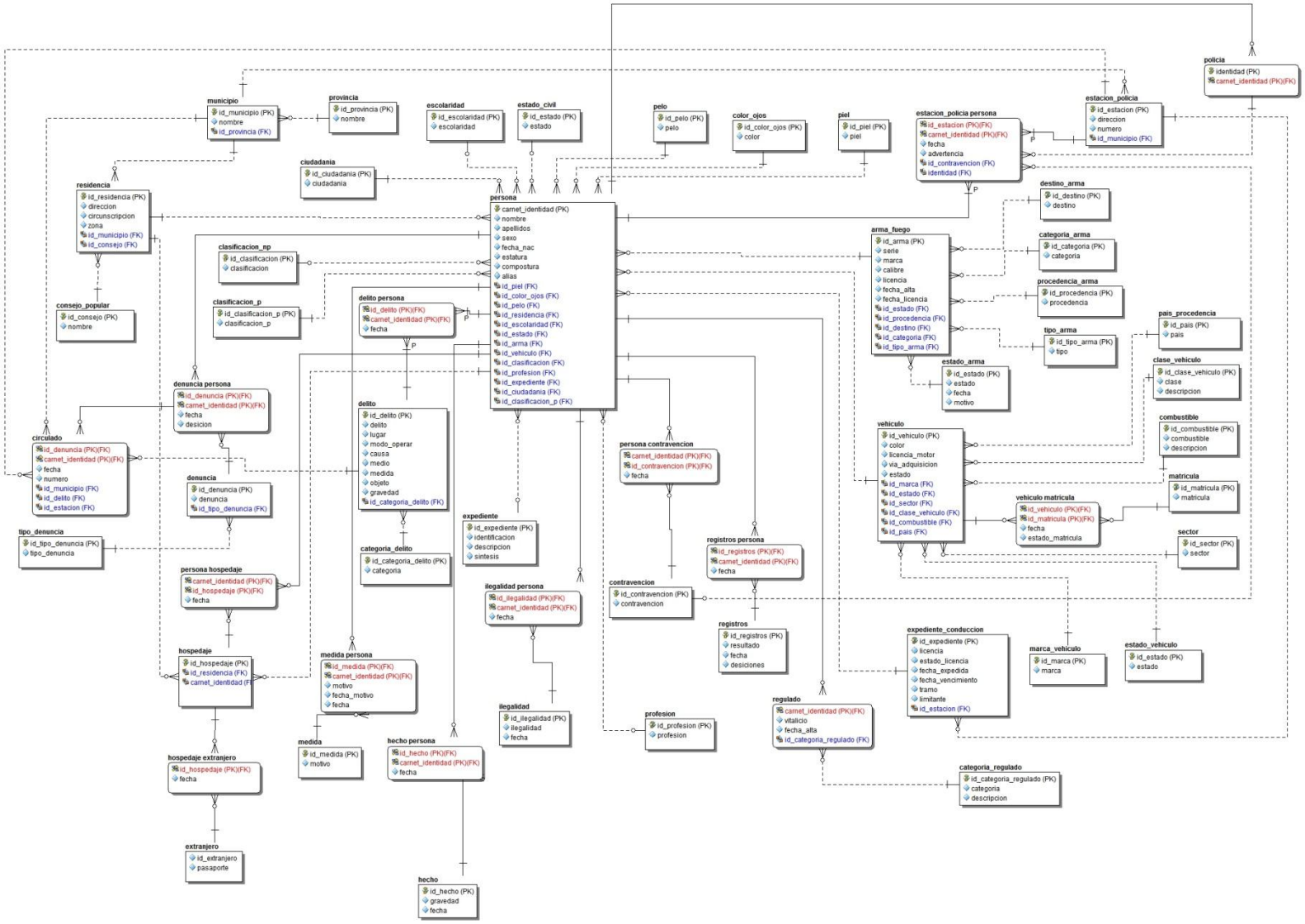
Para realizar la función de staging área se escogió una base de datos relacional diseñada para el gestor de base de datos Oracle. La misma está normalizada en 3ra forma y cuenta con todas las tablas y campos para cargar los datos sin redundancia y manteniendo la integridad.

Con el uso del perfil y mapa de datos se podrá insertar todos los datos de forma que los datos que provienen de fuentes heterogéneas estén integrado de manera uniforme.

La base de datos de staging área propuesta está compuesta por:

- 57 tablas
- 210 columnas
- 56 índices
- 65 llaves foráneas

A partir de esta base datos se realizar el diseño de una base datos dimensional con la arquitectura de estrella o copo de nieve para su carga en un datamart.
A continuación se muestra el modelo de datos de la staging área realizado.


municipio
id_municipio (PK)
nombre
id_provincia (FK)
provincia
id_provincia (PK)
nombre
escolaridad
id_escolaridad (PK)
escolaridad
estado_civil
id_estado (PK)
estado
pelo
id_pelo (PK)
pelo
color_ojos
id_color_ojos (PK)
color
piel
id_piel (PK)
piel
estacion_policia persona
id_estacion (PK)(FK)
carnet_identidad (PK)(FK)
fecha
advertencia
id_contravencion (FK)
identidad (FK)
estacion_policia
id_estacion (PK)
direccion
numero
id_municipio (FK)
policia
identidad (PK)
carnet_identidad (PK)(FK)
ciudadania
id_ciudadania (PK)
ciudadania
residencia
id_residencia (PK)
direccion
circunscripcion
zona
id_municipio (FK)
id_consejo (FK)
consejo_popular
id_consejo (PK)
nombre
persona
carnet_identidad (PK)
nombre
apellidos
sexo
fecha_nac
estatura
compostura
alias
id_piel (FK)
id_color_ojos (FK)
id_pelo (FK)
id_residencia (FK)
id_escolaridad (FK)
id_estado (FK)
id_arma (FK)
id_vehiculo (FK)
id_clasificacion (FK)
id_profesion (FK)
id_expediente (FK)
id_ciudadania (FK)
id_clasificacion_p (FK)
clasificacion_np
id_clasificacion (PK)
clasificacion
clasificacion_p
id_clasificacion_p (PK)
clasificacion_p
delito persona
id_delito (PK)(FK)
carnet_identidad (PK)(FK)
fecha
destino_arma
id_destino (PK)
destino
arma_fuego
id_arma (PK)
serie
marca
calibre
licencia
fecha_alta
fecha_licencia
id_estado (FK)
id_procedencia (FK)
id_destino (FK)
id_categoria (FK)
id_tipo_arma (FK)
categoria_arma
id_categoria (PK)
categoria
procedencia_arma
id_procedencia (PK)
procedencia
tipo_arma
id_tipo_arma (PK)
tipo
pais_procedencia
id_pais (PK)
pais
estado_arma
id_estado (PK)
estado
fecha
motivo
clase_vehiculo
id_clase_vehiculo (PK)
clase
descripcion
combustible
id_combustible (PK)
combustible
descripcion
vehiculo
id_vehiculo (PK)
color
licencia_motor
via_adquisicion
estado
id_marca (FK)
id_estado (FK)
id_sector (FK)
id_clase_vehiculo (FK)
id_combustible (FK)
id_pais (FK)
matricula
id_matricula (PK)
matricula
vehiculo matricula
id_vehiculo (PK)(FK)
id_matricula (PK)(FK)
fecha
estado_matricula
sector
id_sector (PK)
sector
denuncia persona
id_denuncia (PK)(FK)
carnet_identidad (PK)(FK)
fecha
desicion
circulado
id_denuncia (PK)(FK)
carnet_identidad (PK)(FK)
fecha
numero
id_municipio (FK)
id_delito (FK)
id_estacion (FK)
denuncia
id_denuncia (PK)
denuncia
id_tipo_denuncia (FK)
tipo_denuncia
id_tipo_denuncia (PK)
tipo_denuncia
delito
id_delito (PK)
delito
lugar
modo_operar
causa
medio
medida
objeto
gravedad
id_categoria_delito (FK)
categoria_delito
id_categoria_delito (PK)
categoria
expediente
id_expediente (PK)
identificacion
descripcion
sintesis
persona contravencion
carnet_identidad (PK)(FK)
id_contravencion (PK)(FK)
fecha
contravencion
id_contravencion (PK)
contravencion
registros persona
id_registros (PK)(FK)
carnet_identidad (PK)(FK)
fecha
registros
id_registros (PK)
resultado
fecha
desiciones
ilegalidad persona
id_ilegalidad (PK)(FK)
carnet_identidad (PK)(FK)
fecha
ilegalidad
id_ilegalidad (PK)
ilegalidad
fecha
persona hospedaje
carnet_identidad (PK)(FK)
id_hospedaje (PK)(FK)
fecha
hospedaje
id_hospedaje (PK)
id_residencia (FK)
carnet_identidad (F
medida persona
id_medida (PK)(FK)
carnet_identidad (PK)(FK)
motivo
fecha_motivo
fecha
medida
id_medida (PK)
motivo
hecho persona
id_hecho (PK)(FK)
carnet_identidad (PK)(FK)
fecha
hecho
id_hecho (PK)
gravedad
fecha
hospedaje extranjero
id_hospedaje (PK)(FK)
fecha
extranjero
id_extranjero
pasaporte
profesion
id_profesion (PK)
profesion
regulado
carnet_identidad (PK)(FK)
vitalicio
fecha_alta
id_categoria_regulado (FK)
categoria_regulado
id_categoria_regulado (PK)
categoria
descripcion
expediente_conduccion
id_expediente (PK)
licencia
estado_licencia
fecha_expedida
fecha_vencimiento
tramo
limitante
id_estacion (FK)
marca_vehiculo
id_marca (PK)
marca
estado_vehiculo
id_estado (PK)
estado

CONCLUSIONES GENERALES

Se caracterizó el proceso de Extracción, Transformación y Carga (ETL) de las diversas fuentes de datos del delito en el MININT de Guantánamo. La dispersión de los datos en fuentes heterogéneas arrojó la necesidad realizar el mapeo y perfilado de los datos de dicho proceso. Se describió las tecnologías y herramientas existentes para la extracción, transformación y carga de datos. Se caracterizó las fuentes de datos para la realización del ETL.

Para la realización de un exitoso proceso de ETL, se describió los pasos que deben de realizarse en la primera etapa de mismo; y como resultado se obtuvo un perfil y mapa de los datos de las fuentes de datos, relacionadas con el delito en el MININT en Guantánamo. Adicionalmente se diseñó un área de carga intermedia que sirve de base para las siguientes fases del proceso de ETL.

RECOMENDACIONES

A partir de la investigación y desarrollo de la propuesta recomendamos:

- Continuar con el desarrollo del proceso de ETL, aprovechando el empleo de las nuevas tecnologías.

BIBLIOGRAFÍA

1. ▹ *【Oracle ODI 】Información, Reseñas y Precios | 2020 |*. (s. f.). COMPARASOFTWARE. Recuperado 10 de junio de 2020, de https://www.comparasoftware.com/oracle-data-integrator-odi
2. *5.2.1 ALMACENES DE DATOS (DATA WAREHOUSE)—Maria del Socorro Rosas Gaspar*. (s. f.). Recuperado 10 de junio de 2020, de https://sites.google.com/site/mariarosasgaspar/5-2-1-almacenes-de-datos-data-warehouse
3. *Aggregate Profiler User Interface | Download Scientific Diagram*. (s. f.). Recuperado 10 de junio de 2020, de https://www.researchgate.net/figure/Aggregate-Profiler-User-Interface_fig7_221269155
4. Almeida, F. (2017). *Concepts and Fundaments of Data Warehousing and OLAP*.
5. Brachman, R. J., & Anand, T. (1996). The process of knowledge discovery in databases. En *Advances in knowledge discovery and data mining* (pp. 37–57). American Association for Artificial Intelligence.
6. *CloverDX | Solve demanding, real-world data challenges*. (s. f.). Recuperado 10 de junio de 2020, de https://www.cloverdx.com/
7. *CRM On Demand, Soluciones CRM On Demand de Salesforce—Salesforce España*. (s. f.). Recuperado 10 de junio de 2020, de https://www.salesforce.com/es/
8. Curto, J. (2006, noviembre 28). DW: Definiciones de Inmon y Kimball. *Josep Curto*. http://josepcurto.com/2006/11/28/dw-definiciones-de-inmon-y-kimball/

9. *Data profiling, el primer paso en calidad de datos*. (s. f.). Recuperado 10 de junio de 2020, de https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/data-profiling-el-primer-paso-en-calidad-de-datos
10. *DataCleaner: Data Analysis*. (s. f.). Quadient. Recuperado 10 de junio de 2020, de https://po.quadient.com/en/resources/datacleaner-data-analysis
11. *Datamart*. (s. f.). Recuperado 10 de junio de 2020, de https://www.sinnexus.com/business_intelligence/datamart.aspx
12. *Descubrimiento Patrones Desempeño Académico—Libro* | *Procesamiento de datos*. (s. f.). Scribd. Recuperado 10 de junio de 2020, de https://es.scribd.com/document/348548789/DescubrimientoPatronesDesempeno-Academico-Libro
13. *Did you know? Sabía que SAS es número 1 en inteligencia artificial y analítica* | *SAS*. (s. f.). Recuperado 10 de junio de 2020, de https://www.sas.com/es_cl/company-information/discover/ai-analytics-platform.html
14. *El paquete de SSIS*. (s. f.). Recuperado 10 de junio de 2020, de https://support.microsoft.com/es-cr/help/918760/ssis-package-does-not-run-when-called-from-a-sql-server-agent-job-step
15. *ETL de código abierto e integración de datos gratuita: Talend Open Studio*. (s. f.). Recuperado 10 de junio de 2020, de https://es.talend.com/products/talend-open-studio/
16. Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd Edition). The Morgan Kaufmann Series in Data Management Systems. https://www.google.es/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&u

act=8&ved=0ahUKEwi518jO7P3YAhXP21MKHejOBWoQFggsMAA&url=http%3A%2F%2Fmyweb.sabanciuniv.edu%2Frdehkharghani%2Ffiles%2F2016%2F02%2FThe-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf&usg=AOvVaw0ez0loCjWdD0Jp2-2B3bAu

17. *Herramientas Big Data más usadas en la actualidad*. (s. f.). Recuperado 10 de junio de 2020, de https://revistadigital.inesem.es/informatica-y-tics/herramientas-big-data/
18. Inc, C. (s. f.). *What happened to Community Edition*. Recuperado 10 de junio de 2020, de https://www.cloverdx.com/what-happened-to-cloveretl-community-edition
19. Inmon, W. H. (2002). *Building the Data Warehouse,3rd Edition* (3rd ed.). John Wiley & Sons, Inc.
20. Martínez, A. B., Lista, E. A. G., & Flórez, L. C. G. (s. f.). *Técnicas de modelado de procesos de ETL: una revisión de alternativas y su aplicación en un proyecto de desarrollo de una solución de BI*. *18*(1), 8.
21. *Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration | Wiley*. (s. f.). Wiley.Com. Recuperado 10 de junio de 2020, de https://www.wiley.com/en-us/Pentaho+Kettle+Solutions%3A+Building+Open+Source+ETL+Solutions+with+Pentaho+Data+Integration-p-9780470635179
22. *Principales categorías de herramientas ETL - BI Geek Blog*. (s. f.). Recuperado 10 de junio de 2020, de https://blog.bi-geek.com/4-tipos-herramientas-etl/

23. *¿Qué es Pentaho Data Integration (PDI) y para qué sirve?* (s. f.). Recuperado 10 de junio de 2020, de https://www.itop.academy/blog/item/que-es-pentaho-data-integration-pdi-y-para-que-sirve.html
24. *SAS Help Center: About SAS and DataFlux*. (s. f.). Recuperado 10 de junio de 2020, de https://documentation.sas.com/?docsetId=whatsdiff&docsetTarget=n0k0t20kfepbo9n1ktvgchplt1yk.htm&docsetVersion=9.4&locale=en
25. *Soluciones XML, de integración de datos y de desarrollo móvil de Altova*. (s. f.). Recuperado 10 de junio de 2020, de https://www.altova.com/
26. Teiken, Y. (2012). *Automatic Model Driven Analytical Information Systems*. Logos Verlag Berlin GmbH.
27. *Validación de Técnicas de Migración y Herramientas Etcl | Servidor SQL de Microsoft | SQL | Prueba gratuita de 30 días | Scribd*. (s. f.). Recuperado 10 de junio de 2020, de https://es.scribd.com/document/423257075/Validacion-de-Tecnicas-de-Migracion-y-Herramientas-Etcl
28. *What is data profiling and how does it make big data easier?* (s. f.). Recuperado 10 de junio de 2020, de https://www.sas.com/en_us/insights/articles/data-management/what-is-data-profiling-and-how-does-it-make-big-data-easier.html