CADA HECHO DE TU VIDA Cuenta

# INFORME TÉCNICO DE PROCESAMIENTO DE BASES DE DATOS

Mayo, 2020


INEC Buenas cifras, mejores vidas

## Contenido

# INFORME TÉCNICO DE PROCESAMIENTO DE BASES DE DATOS.

## 1. ANTECEDENTES

El Instituto Nacional de Estadística y Censos – INEC viene trabajando en la creación del Sistema Nacional de Registros Administrativos – SNRA con fines estadísticos, mismo que proyecta integrar datos provenientes de distintas fuentes administrativas y aprovecharlos para la generación de estadísticas.

Los registros administrativos que manejan las instituciones públicas (o privadas), constituyen una alternativa con un gran potencial para ampliar y mejorar las estadísticas oficiales del país, esto debido a la captación continua de datos actualizada, veraz y de bajo costo, lo cual es fundamental para satisfacer las necesidades de información de todos los sectores de la sociedad, tanto públicos como privados.

En el informe, se presenta el detalle de los procesos realizados para lograr obtener una base de datos de óptima calidad, con el propósito de utilizar la base de datos para la generación de información estadística."

## 2. OBJETIVO

Informar las etapas de procesamiento de las bases de datos, con propósitos de obtener el Sistema Nacional de Registros Administrativos-SNRA.

## 3. DESARROLLO

En el procesamiento de la base de datos consta de etapas con fin de ir depurando la base de datos, la cuales son:

- Fase de Captación.
- Fase de Perfilamiento.
- Fase de Corrección.
- Fase de Estandarización.
- Fase de Precisión.
- Fase de Coherencia
- Fase de Unicidad.
- Fase de Seudonimización.
- Fase de Integración

### 3.1 CAPTACIÓN

En esta fase se tiene dos plataformas de administración de Datos como son; Hadoop (Clustersg)[1] y Oracle[2], lo que permite administrar, dar un tratamiento y análisis de la

[1] *Apache Hadoop es un framework de software que soporta aplicaciones distribuidas bajo una licencia libre.*
[2] *Es un sistema de gestión de base de datos de tipo objeto-relacional (ORDBMS, por el acrónimo en inglés de Object-Relational Data Base Management System), desarrollado por Oracle Corporation.*

información de las bases de datos internas como externas.

En el Oracle se reciben las bases de datos internas y externas conforme al medio de transferencia (FTP, Vistas Materializadas), acordado previamente con los departamentos o entidades involucradas, una vez que las bases se encuentren en el Oracle, se las organiza con fin de optimizar recursos de memoria y almacenamiento.

Conforme a los requerimientos internos o externos, algunas bases serán seudonimizadas, por motivo que no debe ir datos sensibles (cedula, ruc, nombres, etc.), y puedan los requirentes trabajar sobre la información, este proceso se lo realiza en el Oracle. Una de las actividades adicionales es generar backups de las bases, con procesos automáticos y que son almacenados en un servidor.

En el Clúster se realiza el procesamiento general de las bases de datos, por ser una herramienta de alta prestaciones y procesa grande volúmenes de información. Las bases son subidas por el administrador, dándonos los permisos correspondientes para poder trabajar sobre ella, todo el procesamiento realizado será almacena en el clúster, adicional a ello nos permite realizar una análisis general de las bases de datos.

## 3.2 PERFILAMIENTO

En esta fase se procede a la validación de cada una de las variables con el fin de verificar el estado actual de la base de datos y cuáles serán los procesos posteriores a ejecutar.

Para ejecutar el proceso de validación general de la base de datos, se utiliza UDF-(Funciones definidas por el usuario)[3], esto permite validar la base y posterior a ello dar un diagnóstico. Con esta información proporcionada de la validación, definimos la metodología a usar para continuar con el tratamiento de la base de datos.

En la ejecución del perfilamiento identificamos las variables de la fuente de información que se van evaluar, esto se realiza conforme a una directriz para procesamiento, que se encuentra normada por metodología. Una vez realizado el proceso se realiza los respectivos conteos de los códigos de validación, elaborando un catálogo que ayuda a determinar el estado actual y la calidad de las base de datos.

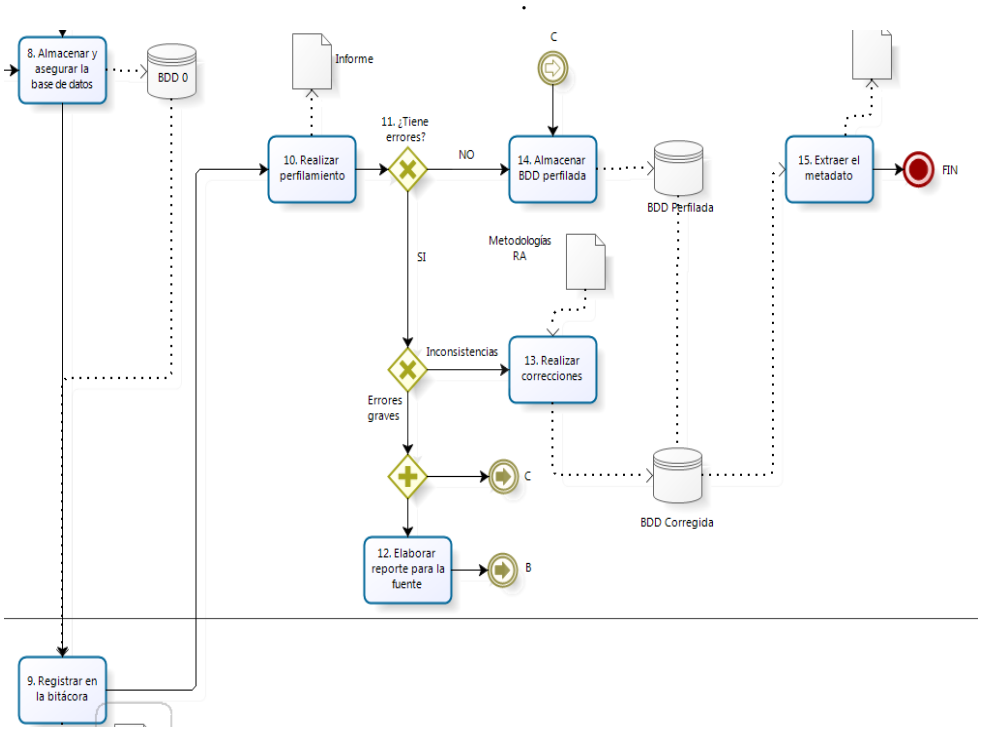
Una calidad de datos se refiere a los procesos, técnicas, algoritmos y operaciones encaminados a mejorar la calidad de los datos existentes, ya que permite acelerar

[3] *UDF: las funciones definidas por el usuario (a menudo abreviado como UDF) permiten codificar su propia lógica de la aplicación para el procesamiento de valores de columna en una consulta en hadoop, desarrollados en lenguaje java.*

crecimiento. La gestión de la calidad de los datos es crítica en un entorno de Data Warehouse[4] y posteriormente analítico porque, cualquier error en calidad de datos, hace que la información extraída posteriormente por el sistema de Business Intelligence[5] pueda conducir a la toma de decisiones erróneas.

Una vez realizado el proceso se llena una ficha de calidad de la base de datos, para determinar el porcentaje de calidad, de cada una variables definidas por la directriz.

**Ilustración 1 Proceso de Perfilamiento**




La base perfilada es almacenada en el clúster y posteriormente nos ayudara a verificar las inconsistencias del mismo, luego de ello se procederá a realizar la corrección de cada una de las variables. Para realizar poder obtener el diagnostico verificamos en los códigos de la validación que son:

Validación cedula:

| Descripción | Código |
|---|---|

[4] *Datawarehouse es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de pespectivas y con grandes velocidades de respuesta.*
[5] *Business Intelligence es la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios.*

| | |
|---|---|
| Cédula correcta, paso el dígito verificador | CC00 |
| Cédula con valor en nulo o vacíos | CC01 |
| Cédula con mayor de 10 dígitos | CC02 |
| Cédula con 9 dígitos (Se adiciona un cero adelante) | CC03 |
| Cédula con menos de 9 dígitos | CC04 |
| Casos nuevos | CC05 |
| Contiene caracteres especiales y/o letras | CC06 |
| Contiene espacios en blanco | CC07 |
| Cédulas inician con 61 | CC08 |
| Dígito verificador incorrecto | CC09 |

Validación RUC:

| Descripción | Código |
|---|---|
| RUC correcto. | RR00 |
| RUC con valor nulo o vacío. | RR01 |
| RUC mayor que 13 dígitos. | RR02 |
| RUC igual a 12 dígitos. | RR03 |
| RUC menor a 12 dígitos. | RR04 |
| Casos nuevos. | RR05 |
| RUC con caracteres especiales o letras. | RR06 |
| RUC con espacios en blanco. | RR07 |
| Tres últimos dígitos incorrectos. | RR08 |
| Dígito verificador incorrecto. | RR09 |
| Los 2 primeros dígitos igual a 61. | RR10 |
| Segundo parámetro fuera de rango | RR11 |

Validación de Fecha:

| Descripción | Código |
|---|---|
| Fecha correcta | FC00 |
| Si nulo o vacío | FC01 |
| Igual a 7 caracteres | FC02 |
| Día fuera de rango | FC03 |
| Mes fuera de rango | FC04 |
| Año fuera de rango | FC05 |
| Sólo año | FC06 |
| Formato ingresado incorrecto | FC08 |
| Caracteres especiales | FC09 |
| Fecha mayor a 8 dígitos | FC10 |
| Fecha menor a 7 dígitos | FC11 |
| Error Desconocido | FC12 |
| Fecha mayor a la actual | FC13 |

Validación de Números Decimales:

| Descripción | Código |
|---|---|
| Si es correcta. | DC00 |

| | |
|---|---|
| Si es nulo y cadena vacía. | DC01 |
| Si tiene espacios en blanco. | DC02 |
| Si tiene caracteres especiales y Contiene letras. | DC03 |
| Formato incorrecto. | DC04 |
| Error Desconocido | DC05 |

Validación de Números Enteros:

| Descripción | Código |
|---|---|
| Si es correcta. | NM00 |
| Si es nulo y cadena vacía. | NM01 |
| Si tiene espacios en blanco. | NM02 |
| Si tiene caracteres especiales. | NM03 |
| Error desconocido | NM04 |

Validación de Texto o Cadena:

| Descripción | Código |
|---|---|
| Si es correcta. | AA00 |
| Si es nulo y cadena vacía | AA01 |
| Si tiene caracteres especiales. | AA03 |
| Otro error | AA04 |
| Segundo parámetro fuera de rango | AA05 |

Validación de letras y números (alfanumérico):

| Descripción | Código |
|---|---|
| Si es correcta. | NL00 |
| Si es nulo y cadena vacía. | NL01 |
| Si tiene espacios en blanco. | NL02 |
| Si tiene caracteres especiales. | NL03 |
| Sólo letras. | NL04 |
| Sólo números. | NL05 |

## 3.3 CORRECCIÓN

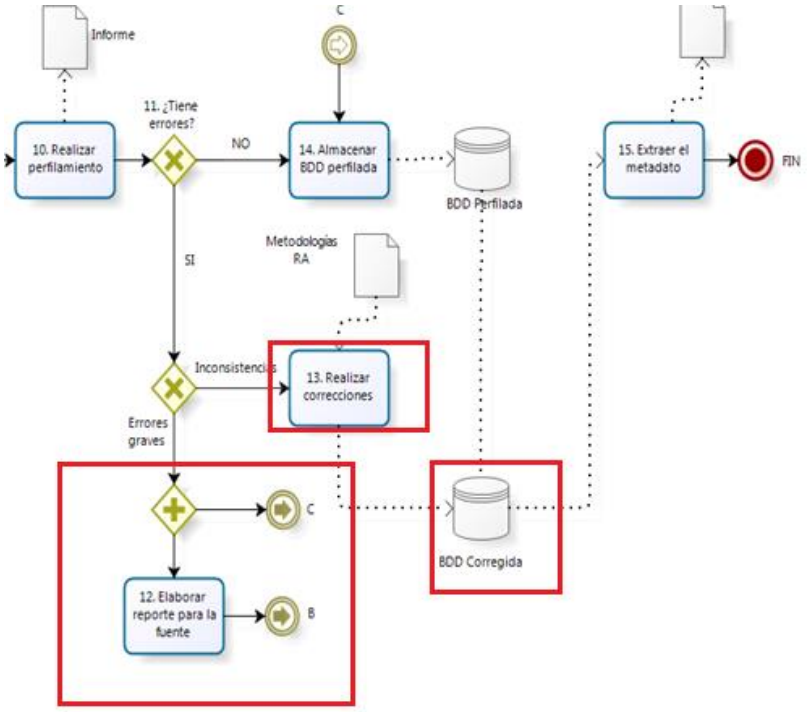
En la etapa de corrección de los datos de variables, esto se procede mediante los UDF (funciones definidas por los usuarios), conforme a la validación realizada se le envía los mismos parámetros para la corrección.

Se lo realiza conforme a las inconsistencias existentes, una vez realizado la corrección se verifica que las variables se encuentren correctamente corregidas y los casos específicos (variables que no pueden ser corregidas por error desde la fuente), serán reportados por el área de metodología a las instituciones proveedoras de la información, con fines de retroalimentación con las bases de datos.

Con la base corregida se obtiene un catálogo de las variables con fin a determinar la

calidad actual de las bases de datos.

**Ilustración 2 Proceso de Corrección**




La limpieza de las variables se encuentra previamente definida y revisada metodológicamente en la directriz y clasificado conforme al código de validación, se realiza lo siguiente:

En caso de corrección de cedula:

| Código | Corrección aplicada |
|---|---|
| CC00 | Retorna la misma cédula. |
| CC01 | Retorna cadena vacía. |
| CC02 | Cédula con los mismos dígitos de entrada. |
| CC03 | Se corrige con un 0 delante siempre y cuando pase el algoritmo. |
| CC04 | Cédula con los mismos dígitos de entrada. |
| CC05 | Error desconocido |
| CC06 | Elimina caracteres y letras, reemplaza O por 0 y S por 5. |
| CC07 | Elimina los espacios en blanco. |
| CC08 | No aplica corrección. |
| CC09 | Cédula con los mismos dígitos de entrada. |

Corrección de RUC:

| Código | Corrección aplicada |
|---|---|
| RR00 | Retorna el mismo RUC |

| | |
|---|---|
| RR01 | Retorna cadena vacía |
| RR02 | RUC con los mismos dígitos de entrada |
| RR03 | Corrige aumentando un cero al inicio, solo cuando el RUC sea correcto. |
| RR04 | RUC con los mismo dígitos de entrada |
| RR05 | Error desconocido |
| RR06 | Elimina caracteres y letras, reemplaza O por 0 y S por 5. |
| RR07 | Elimina los espacios en blanco. |
| RR08 | Retorna el mismo RUC. |
| RR09 | RUC con los mismos dígitos de entrada |
| RR10 | No aplica corrección |
| RR11 | Devuelve el mismo valor que ingreso |

Corrección de Fecha, se considera el formato de la fecha que llega y se le envié como parámetro para realizar la corrección:

| Código | Corrección aplicada |
|---|---|
| FC00 | Retorna con el formato año-mes-día |
| FC01 | Retorna fecha por default 0001-01-01 |
| FC02 | Se corrige con un 0 delante cuando la fecha es válida y el formato de la fecha:<br>Formato fecha 1 y 2 se añade un cero adelante.<br>Formato fecha 3 se añade un cero en el antepenúltimo digito (ej: 2016063 se corrige 20160603) |
| FC03 | Si día es mayor a número de días del mes en curso (28, 29, 30 ó 31) ó menor a 1, entonces día igual a 15. |
| FC04 | Si mes es mayor a 12 ó menor a 1, entonces mes igual a 06. |
| FC05 | Si el año de la fecha es fuera de rango menor a 1800 o mayor a la fecha actual, se deja (0001-01-01) (dependiendo del tercer parámetro: fec_futuro). |
| FC06 | Si fecha tiene de longitud tamaño 4, si el año es fuera de rango: menor a 1800 o mayor a la fecha actual, se deja 0001-01-01 (dependiendo del tercer parámetro: fec_futuro).<br>Cuando el año este dentro del rango se completara la fecha imputando el mes y año a mitad del año con el ultimo día (06-30). |
| FC08 | Si el formato de la fecha (1, 2, 3) ingresado incorrecto, devuelve el mismo valor que ingresó (dependiendo del segundo parámetro: formato). |
| FC09 | Elimina caracteres especiales |
| FC10 | Devuelve la misma cadena |
| FC11 | Devuelve la misma cadena |
| FC12 | Error no catalogado |
| FC13 | Devuelve 0001-01-01 |

En caso de corrección de números decimales:

| Código | Corrección aplicada |
|---|---|
| DC00 | Retorna el mismo número |
| DC01 | Retorna cadena vacía |
| DC02 | Elimina los espacios en blanco |
| DC03 | Elimina los caracteres especiales y letras |
| DC04 | Retorna el mismo número |
| DC05 | Error no catalogado |

Corrección de números enteros:

| Código | Corrección aplicada |
|---|---|
| NM00 | Retorna el mismo número |
| NM01 | Retorna cadena vacía |
| NM02 | Elimina los espacios en blanco |
| NM03 | Elimina los caracteres especiales |
| NM04 | Error no catalogado |

Corrección de letras o cadena, se le verifica el tipo de corrección que se desea realizar, si envié como parámetro 1 valida solo cadena y 0 valida texto y números:

| Código | Corrección aplicada |
|---|---|
| AA00 | Retorna la misma cadena |
| AA01 | Retorna cadena vacía |
| AA03 | Elimina los caracteres especiales, Elimina las tildes a las vocales, Transforma las letras a mayúsculas<br>Deja sólo un espacio entre palabra y borra espacios al inicio y final. |
| AA04 | Retorna la misma cadena |
| AA05 | Devolver la misma cadena que ingreso |

Corrección de letras y números (alfanumérico):

| Código | Corrección aplicada |
|---|---|
| NL00 | Retorna el mismo número o cadena |
| NL01 | Retorna campo vacía |
| NL02 | Elimina los espacios en blanco y deja sin espacios entre cadenas. |
| NL03 | Elimina los caracteres especiales |
| NL04 | Retorna el mismo carácter |
| NL05 | Retorna el mismo número |

## 3.4 ESTANDARIZACIÓN

Para poder realizar la estandarización se utiliza la plataforma METADEC[6], que tiene como objetivo organizar diferentes objetos de información estadística dentro de una estructura común para referenciarlos hacia estándares de almacenamiento y transferencia de metadatos, facilitar la homologación de las categorías y clasificaciones de variables, estandarizar la definición de variables y validaciones para captura de información y reutilizar objetos de información para construcción de cuestionarios de captura.

Para la estandarización de las variables de una base de datos, el equipo de metodología realiza el proceso de elaborar los catálogos de correspondencia de las variables categóricas, los mismos en el METADEC, para poder realizar un cruce y

[6] *METADEC es un sistema para gestión de metadatos basado en terminología utilizada en estándares internacionales. Permite la estandarización de datos provenientes de diferentes fuentes. Además sirve para la construcción automática de cuestionarios para captura.*

normar las variables se procede a realizar una conexión con el clúster mediante el uso de UDF (Funciones definidas por el usuario), de esta manera tendrán una correspondencia con nuestros registros.

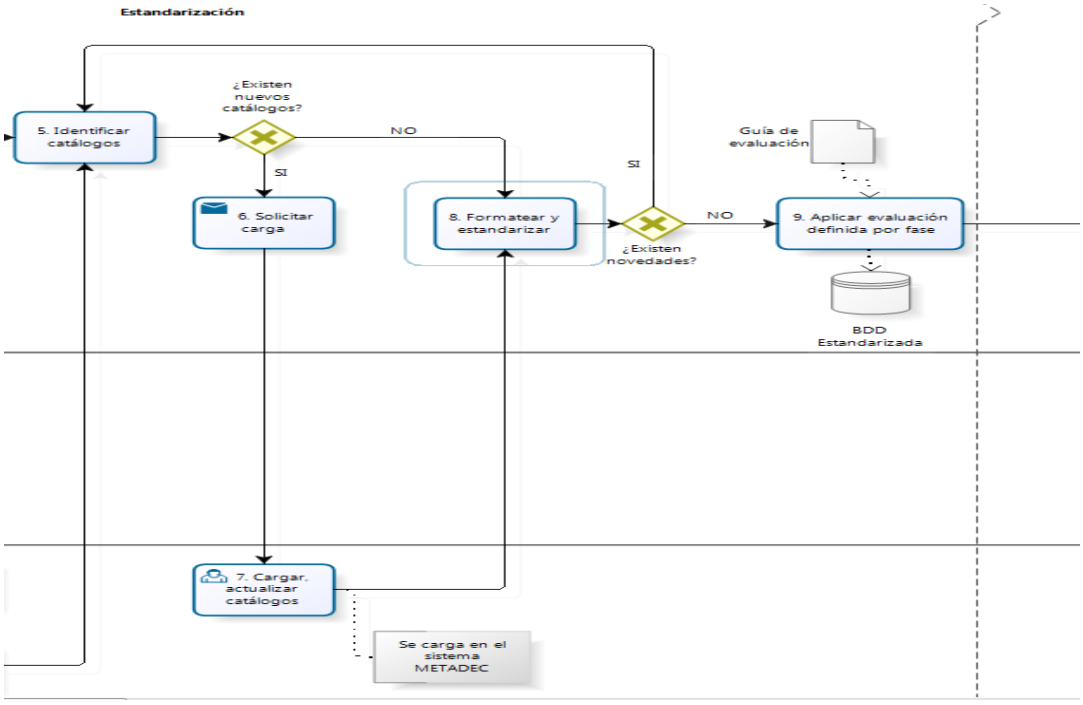
Todas las categorizaciones realizadas se las identifica con la nomenclatura for ejemplo (for_lugar_nacimiento), de esta forma se puede verificar si no existen errores en el catálogo subido en el METADEC.

Para poder identificar los posibles errores en el catálogo, nos ayudamos con las siguientes nomenclaturas:

- NE: No existe o no encontrado valor (no se encuentra en el catálogo).
- NC: No tiene catálogo.
- NA: No aplica (Error desconocido).

Cuando se detecta un error de estos, se procede a la revisión de los catálogos subidos en la meta data junto con los datos con el fin de corregir y proceder nuevamente a estandarizar. Una catalogo no correctamente definido ocasiona retrasos debido a la revisión manual de los errores y retroalimentar al equipo de metodología para las correcciones del caso.

**Ilustración 3 Proceso de Estandarización**




## 3.5 PRECISIÓN

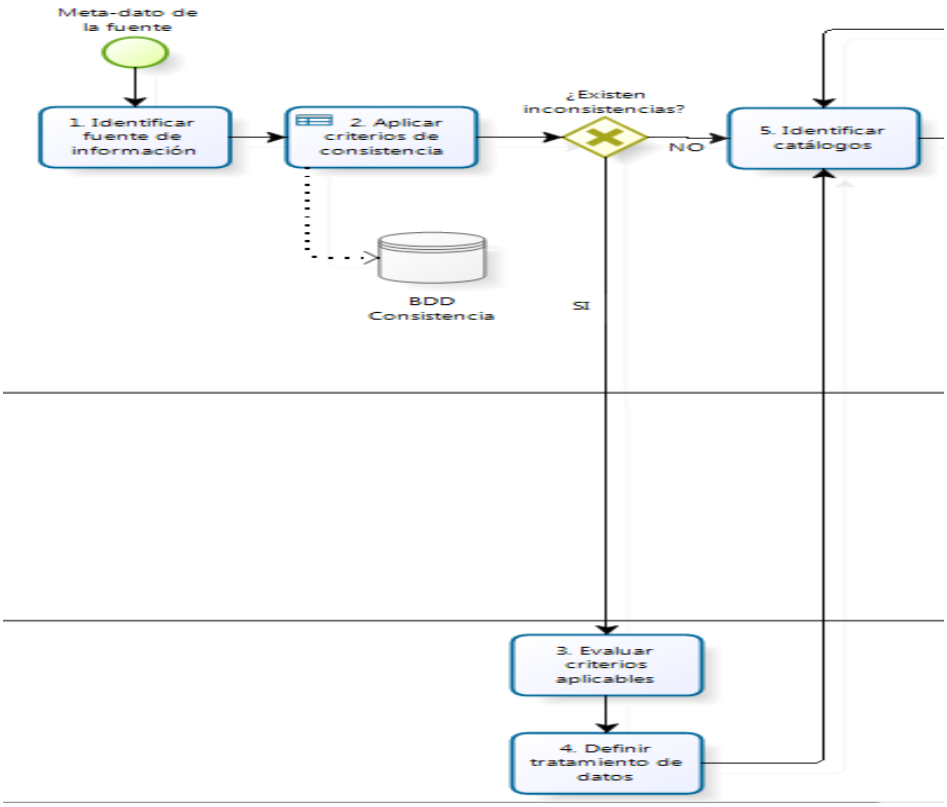
En la etapa de precisión, se realiza el proceso de la recuperación de las variables de identificación[7] (cedula, ruc, etc.), los registros serán recuperados por medio del info-digital de manera manual, por el equipo de metodología. Otro proceso que se realiza es la similitud de nombres[8], para obtener la variable de identificación conforme al mayor porcentaje de similitud de la variable, por lo general se considera desde el 0,85% en adelante.

Previo a esto se realiza cruces con la base de datos mandatorio con fines de recuperar la variable de identificación, conforme a la metodología establecida obteniendo datos más verídicos y confiables. Los datos recuperados son identificados mediante el campo cal_id (calidad de la variable de identificación).

Concluido la recuperación de variables de identificación se procede a obtener un catálogo de todas las variables categóricas[9] .

**Ilustración 4 Proceso de Precisión**

[7] *Variable o conjunto de variables, que permiten identificar a un objeto o conjunto de objetos en un registro estadístico. Por ejemplo: número cédula, Registro Único de Contribuyente – RUC, etc.*
[8] *Son métodos y sus algoritmos de coincidencia correspondientes, son parte del criterio de coincidencia con reglas establecidas. Ayudan a determinar cómo se compara un campo específico en un registro con el mismo campo en otro registro y si los campos se consideran coincidencias o similares.*
[9] *Las variables categóricas también se denominan variables cualitativas o variables de atributos. Los valores de una variable categórica se pueden colocar en un número contable de categorías o grupos diferentes. Los datos categóricos pueden tener o no algún orden lógico.*




## 3.6 COHERENCIA

En la etapa de coherencia, se realizan validaciones entre variables de un mismo registro administrativo, de tal forma que permita verificar la coherencia de los datos.

## 3.7 UNICIDAD

En la etapa de unicidad, se revisan que los registros sean únicos a nivel de unidad de análisis; es decir, identificar las unidades duplicadas. Para identificar duplicados se requiere de un análisis detallado o individual para cada registro administrativo, ya que en muchos casos se consideraran más de una variable del registro, sobre todo los que registran datos de diferente periodo.

## 3.8 SEUDONIMIZACION

La protección de la confidencialidad de identidad de los datos provenientes de fuentes administrativas implican dos fases: la Seudonimización y la anonimización.

La seudonimización corresponde al reemplazo de los datos de identificación por un código de identificación (aleatorio) generado en el INEC, código que debe ser el mismo en todas las bases de datos a fin de identificar y vincularlos entre las distintas

bases de datos. Mientras que en la anonimización se debe eliminar los códigos generados en la seudonimización, además, se debe analizar cuidadosamente el riesgo de revelación de datos confidenciales en los tabulados que se pueden generar.

El proceso de la seudonimización tiene como entrada la base de datos con todas las variables provenientes de una fuente administrativa y pasarlo por una BDD donde se almacenan las identificaciones individuales y se realiza el cruce con las correspondencias (código INEC).

Los casos de seudonimización que tienen el código inicial 88 son casos específicos como en Afiliados que tiene números de RUC que no constan en la base de contribuyentes, por lo cual se procedido a colocar una codificación diferente y se puede identificar con facilidad los caso.

Para proceder a realizar la seudonimización se ha considerado los siguientes criterios para cada caso:

a) Criterios de validación de la Función (SEUDONIMIZAR_AFILIADOS):
    - Comparar la cantidad de registros entre la tabla original y la tabla generada.
    - Comparar la cantidad de registros únicos entre la variable cedula perteneciente a la tabla original y la variable cod_inec_ci perteneciente a la tabla generada.
    - Comparar la cantidad de registros únicos entre la variable ruc perteneciente a la tabla original y la variable cod_inec_ruc perteneciente a la tabla generada.
    - Verificar que exista un solo cod_inec_ci para cada número de Cédula, ya que no deben existir duplicados.
    - Verificación de que existan un solo Cod_Inec_Ruc para cada número de Ruc. No deben existir duplicados.
    - Se debe verificar que para los números de Ruc, en donde, los 10 primeros dígitos son iguales y los 3 últimos son distintos, debe corresponder un mismo Cod_Inec_Ruc.
    - Verificación de que para la variable Cod_Inec_Ci no deben existir valores nulos.
    - Verificación de que para la variable Cod_Inec_Ruc no deben existir valores nulos.

b) Criterios de validación de la Función (SEUDONIMIZAR_CEDULA):

- Comparar la cantidad de registros entre la tabla original y la tabla generada.
- Comparar la cantidad de registros únicos entre la variable cedula perteneciente a la tabla original y la variable cod_inec_ci perteneciente a la tabla generada.
- Verificar que exista un solo cod_inec_ci para cada número de Cédula, ya que no deben existir duplicados.
- Verificación de que para la variable Cod_Inec_Ci no deben existir valores nulos.

c) Criterios de validación de la Función (SEUDONIMIZAR_RUC):

- Comparar la cantidad de registros entre la tabla original y la tabla generada.
- Comparar la cantidad de registros únicos entre la variable ruc perteneciente a la tabla original y la variable cod_inec_ruc perteneciente a la tabla generada.
- Verificación de que existan un solo cod_inec_ruc para cada número de ruc. No deben existir duplicados.
- Se debe verificar que para los números de ruc, en donde, los 10 primeros dígitos son iguales y los 3 últimos son distintos, debe corresponder un mismo cod_inec_ruc.
- Verificación de que para la variable cod_inec_ruc no deben existir valores nulos.

## 3.9 INTEGRACIÓN

La integración de datos se ejecuta con dos propósitos, uno para la construcción de tablas históricas y otra para la generación de registros temáticos.

| | NOMBRE | CARGO | FIRMA |
|---|---|---|---|
| Elaborado por: | Stalyn Flores | Analista de Registros Administrativos | |
| Revisado y por: | Ángel Chiluisa | Responsable de Gestión del Sistema de Registros Administrativos | |
| Aprobado por: | David Caín | Director (E) de Registros Administrativos | |

# CADA HECHO DE TU VIDA *Cuenta*

@ecuadorencifras

@InecEcuador

t.me/ecuadorencifras

INEC/Ecuador

INECEcuador

INEC Ecuador