Plan de Validación y Tabulación

Directorio de Empresas y Establecimientos

instituto nacional de estadística y censos
www.ecuadorencifras.gob.ec

Contenido

Introducción

El Directorio de Empresas y Establecimientos maneja una gran cantidad de variables, las cuales deben ser validadas, de tal manera que se tenga información confiable y veraz al momento de elaborar las publicaciones anuales de un nuevo Directorio.

El software que sirve como herramienta para la Validación de variables, se emplea para la realización validaciones automáticas, pero también dentro del proceso es indispensable la realización de validaciones manuales, todo esto con el objetivo de obtener información de mejor calidad.

Para la validación, es importante considerar que existen variables que dependen de otras para ser validadas y otras que son independientes, es decir, que no requieren de ninguna variable adicional para su validación.

El presente documento contiene el plan de validación y tabulación de las variables que existen en el DIEE, basado en la experiencia obtenida en cada publicación, también se incluyen documentos de la limpieza de datos detallados en una matriz a tomar en cuenta al momento de proceder con el arreglo de los mismos, análisis de inconsistencias en la información o algún tipo de novedad adicional que se dé en el manejo de la base de datos.

Variables involucradas

Existe una gran cantidad de variables, estas después de su validación deben ser subidas a la base de datos del Directorio de Empresas y Establecimientos, por lo que, se ha creado una matriz de priorización de validación de datos por cada variable de análisis.

Matriz de prioridad:

Prioridad	Descripción
1	Alta
2	Media
3	Baja

Valoración de la matriz de prioridad:

Alta: Variables que están directamente ligadas a la Publicación

Media: Variables que no están en la Publicación pero si tienen relación directa con la publicadas.

Baja: Variables que no tienen relación con la Publicación.

Plan de Validación y Tabulación u_legal (Variables Independientes)

# CRUCE	VARIABLE INICIAL		Validación 1	Validación 2	Validación 3	Prioridad	Detalle
1	**Numero Ruc**		=13 dígitos	Validación del RUC (ANEXO 1)	---	1	Validación Automática
2	**Fecha Desde (FECHA APERTURA)**		8 números consecutivos. En formato **aaaammdd**	---	---	3	Validación Automática
3	**Expediente**		Personas Naturales no tienen expediente.	---	---	3	Validación Automática
4	**Clase Contribuyente**	**Desconocido**	Verificar coincidencia con la fuente	Reportar los casos desconocidos	---	2	Validación Automática
		Especial	Verificar coincidencia con la fuente	---	---	2	Validación Automática
		Otros	Verificar coincidencia con la fuente	---	---	2	Validación Automática
		RISE	Verificar coincidencia con la fuente	---	---	2	Validación Automática
5	**Forma Jurídica**	**Desconocido**	Consultar a la fuente	Reportar los casos desconocidos	---	2	Validación Automática
		Personal Natural	3er digito RUC < 6 PERSONA NATURAL	---	---	2	Validación Automática
		Sociedades	3er digito RUC ≥ 6 EMPRESAS	---	---	2	Validación Automática

Plan de Validación y Tabulación u_legal (Variables Dependientes)

# CRUCE	VARIABLE INICIAL		VARIABLE DE COMPARABILIDAD 1	VARIABLE DE COMPARABILIDAD 2	VARIABLE DE COMPARABILIDAD 3	Validación 1	Validación 2	Validación 3	Prioridad	Detalle
7	**Unidad Legal Estado**	**Desconocido**	Fuente	---	---	Consultar a la Fuente	---	---	1	Reportar los casos desconocidos (Validación Automática)
		Activa	Fecha de inscripción	Fecha de inicio actividad	Fecha de cese actividad	Debe existir: fecha de inscripción y fecha de inicio de actividad	Si ha cerrado debe existir: fecha de cese de actividad	Si ha cerrado debe existir: fecha de reinicio de actividad	1	Validación Automática
		Pasiva	Fecha de inscripción	Fecha de inicio actividad	Fecha de cese actividad	Debe existir: fecha de inscripción	Debe existir: fecha de inicio de actividad	Debe existir: fecha de cese de actividad.	1	Validación Automática
		Suspensión definitiva	Fecha de inscripción	Fecha de inicio actividad	Fecha de cese actividad	Debe existir: fecha de inscripción	Debe existir: fecha de inicio de actividad	Debe existir: fecha de cese de actividad.	1	Validación Automática
		Cerrado	Fecha de inscripción	Fecha de inicio actividad	Fecha de cese actividad	Debe existir: fecha de inscripción	Debe existir: fecha de inicio de actividad	Debe existir: fecha de cese de actividad.	1	Validación Automática
		Por Revisar	Fuente	---	---	Consultar a la Fuente	---	---	1	Reportar los casos desconocidos (Validación Automática)
8	**Obligado a llevar Contabilidad**		Desconocido	---	---	Reportar los casos desconocidos	---	---	2	Validación Automática
9	**Razón Social**		RUC	---	---	(3er digito RUC < 6) =PERSONA NATURAL	(3er digito RUC ≥ 6) = EMPRESAS	---	1	Validación Automática

Plan de Validación y Tabulación Empresa (Variables Independientes)

# CRUCE	VARIABLE INICIAL		Validación 1	Validación 2	Validación 3	Prioridad	Detalle
1	**Nombre Comercial**		Validar vs la Fuente SRI	---	---	1	Validación automática
2	**Actividad Comercio Exterior**	**Mercado Interno**	Debe tener valor codificado (00 ó 99)	---	---	3	Validación automática
		Importador	Debe tener valor codificado (01)	---	---	3	Validación automática
		Exportador	Debe tener valor codificado (02)	---	---	3	Validación automática
		Importador / Exportador	Debe tener valor codificado (03)	---	---	3	Validación automática
3	**Número de Unidad Local**		Valores >= 0	---	---	2	Validación automática
4	**Actividad Económica**		Descripción = Código CIIU 4	---	---	1	Validación automática

Plan de Validación y Tabulación Empresa (Variables Dependientes)

# CRUCE	VARIABLE INICIAL	VARIABLE DE COMPARABILIDAD 1	VARIABLE DE COMPARABILIDAD 2	VARIABLE DE COMPARABILIDAD 3	Validación 1	Validación 2	Validación 3	Validación 4	Prioridad	Detalle
5	**Fecha de inscripción**	Fecha de inicio de actividad	---	---	Debe existir Fecha de inicio.	8 números consecutivos. En formato **aaaammdd**		---	1	Validación automática
6	**Fecha de inicio de actividad**	Fecha de inscripción	---	---	Debe existir Fecha de inscripción	8 números consecutivos. En formato **aaaammdd**		---	1	Validación automática
7	**Fecha de cese de actividad**	Fecha de inscripción	Fecha de inicio de actividad	---	8 números consecutivos. En formato **aaaammdd**			---	1	Validación automática
8	**Fecha de reinicio de actividad**	Fecha de inscripción	Fecha de inicio de actividad	Fecha de cese de actividad	8 números consecutivos. En formato **aaaammdd**			---	1	Validación automática
9	**Fecha de actualización**	---	---	---	8 números consecutivos. En formato **aaaammdd**		---	---	1	Validación automática

Plan de Validación y Tabulación u_local (Variables Independientes)

# CRUCE	VARIABLE INICIAL	Validación 1	Validación 2	Validación 3	Prioridad	Detalle
1	**Numero Ruc**	= 13 dígitos	Validación del RUC (ANEXO 1)	---	1	Validación Automática
2	**Número de unidad local**	Al menos una unidad local registrada.	No pueden existir unidades locales con el mismo número de unidad local.	Validar que no exista el número cero en el identificador.	1	Validación Automática
3	**Nombre Unidad local**	> 3 dígitos	---	---	2	Validación Automática
4	**Actividad Económica**	Descripción = Codigo CIIU 4		---	1	Validación Automática

Plan de Validación y Tabulación u_local (Variables Dependientes)

# CRUCE	VARIABLE INICIAL		VARIABLE DE COMPARABILIDAD 1	VARIABLE DE COMPARABILIDAD 2	VARIABLE DE COMPARABILIDAD 3	Validación 1	Validación 2	Validación 3	Validación 4	Prioridad	Detalle
5	**Fecha de cierre**		Fecha desde	Fecha de apertura	---	8 números consecutivos. En formato **aaaammdd**	---	---	---	3	Validación Automática
11	**Fecha de apertura**		Fecha de inicio de actividad	Fecha de inscripción	---	8 números consecutivos. En formato **aaaammdd**	---	---	---	3	Validación Automática
12	**Fecha de inscripción**		Fecha de inicio de actividad	---	---	8 números consecutivos. En formato **aaaammdd**	---	---	---	3	Validación Automática
13	**Fecha de actualización**		---	---	---	8 números consecutivos. En formato **aaaammdd**	---	---	---	3	Validación Automática
14	**Fecha de reinicio de actividad**		Fecha de inscripción	Fecha de inicio de actividad	Fecha de cese de actividad	En caso de que exista un campo vacío, se coloca (-1).	8 números consecutivos. En formato **aaaammdd**	---	---	3	Debe existir sólo para los registros que presenten fecha de cese Validación Automática
15	**Unidad local estado**	**1 Abierta**	Fecha de apertura	---	---	Existir fecha apertura	Si ha cerrado debe existir: fecha de cese de actividad	Si ha abierto debe existir: fecha de reinicio de actividad	Po lo menos una U legal abierta	3	Validación Automática
		4 Cerrada (Suspensión definitiva de sus actividades)	Fecha de inicio de actividad	Fecha de cierre	---	Existir: fecha de apertura	Existir: fecha de cierre	---	---	1	Validación Automática

17	**Unidad local tipo**	**Desconocido**	Fuente	---	---	Consultar a la Fuente	---	---	---	1	Reportar los casos desconocidos (Validación Automática)
		Matriz	Número de unidades locales	---	---	Una sola MATRIZ	---	---	---	1	Validación Automática
		Auxiliar	Número de unidades locales	---	---	Número de Unidad Local diferente	---	---	---	1	Validación Automática

Plan de Validación y Tabulación Otras Variables

# CRUCE	VARIABLE INICIAL	VARIABLE DE COMPARABILIDAD 1	VARIABLE DE COMPARABILIDAD 2	VARIABLE DE COMPARABILIDAD 3	Validación 1	Validación 2	Prioridad	Detalle
1	**Teléfono**	Provincia			El código telefónico debe corresponder a su respectiva provincia. (EMPRESA)		1	Validación Automática
2	**Empleados**	Empleados Hombres afiliados	Empleados Mujeres afiliadas	Total empleados afiliados y afiliadas	Empresas ΣPA = Σ (PAH + PAM) (PA personal afiliado)	Comprobar en cuanto a unidades locales que la suma de sus empleados sea igual a la que reporta la empresa.	1	Validación Automática
3	**Remuneraciones**	Remuneraciones Hombres	Remuneraciones Mujeres	Total remuneraciones	Empresas ΣR = Σ (RH + RM) (R remuneraciones)	No deben existir valores negativos	1	Validación Automática
4	**Ventas**	Empresas del sector Publico	Empresas Pasivas		Sector público no debe tener ventas (excepto empresas públicas).	Empresas pasivas: Ventas = 0	1	Validación Automática
5	**Ubicación geográfica**	Provincia	Cantón	Parroquia	Si está incompleta, tomar geografía del establecimiento matriz, para empresas con establecimiento único.	Los casos que no se pudo recuperar, entregar un listado de los casos restantes al equipo de Análisis.	1	Validación Automática
6	**Forma institucional**	Forma institucional	Unidad legal tipo	Obligado a llevar contabilidad	Si ulegal_tipo=1 & forma_inst=7 & obligado='N' → forma_inst=2	Si ulegal_tipo=1 & forma_inst=7 & obligado='S' → forma_inst=3	1	Validación Automática
7	**Clase contribuyente**	Clase contriubyente	Declaración RISE	Forma institucional	Si (clase_contrib='RIS' & forma_inst<>1) || RISE=1 → clase_contrib = 'OTR'	Si clase_contrib='OTR' & RISE=1 & forma_inst=1 → clase_contrib = 'RIS'	1	Validación Automática

8	**Estratos ventas y empleo**	Estrato ventas o empleo	Ventas o empleo		Si Tiene ventas>0 debe tener estrato ventas. Si tiene empleo debe tener estrato empleo		1	Validación Automática
9	**Tamaño**	Estrato ventas o empleo			Si Tiene estrato de ventas, debe tener tamaño ó si tiene estrato empleo debe tener tamaño		1	Validación Automática
10	**Situación**	Tamaño			Si tiene tamaño debe tener situación		1	Validación Automática

Valores por defecto y Tratamiento a los datos que presenten inconsistencias

Después del proceso de limpieza, existirán datos que se borren de la base y muchos campos van a quedar en blanco, para estos casos tenemos valores por defecto que vendrán a llenar estos espacios en blanco.

Pero también tendremos variables que si presentan inconsistencias no se deberá incluir al registro dentro de la Base de Datos.

Variable	Inconsistencia	Valor por defecto	Nota
Unidad Legal			
Numero Ruc	No presenta mínimo 13 dígitos	No aplica	El registro no sube y reportar al DIEE
Razón Social	Si la razón social tiene una longitud menor a 3	No aplica	El registro no sube y reportar al DIEE
Expediente	En caso de no existir un valor, se coloca su valor por defecto.	-1	
Acto Jurídico	En caso de no existir un valor, se coloca su valor catalogado	Desconocido	Reportar los casos al DIEE
Clase Contribuyente	En caso de no existir un valor, se coloca su valor catalogado	Desconocido	Reportar los casos al DIEE
Estado Funcionamiento	En caso de no existir un valor, se coloca su valor catalogado	Desconocido	Reportar los casos al DIEE
Forma Jurídica	En caso de no existir un valor, se coloca su valor catalogado	Desconocido	Reportar los casos al DIEE
Unidad Legal Estado	En caso de no existir un valor, se coloca su valor catalogado	Desconocido	Reportar los casos al DIEE
Obligado a llevar Contabilidad	En caso de no existir un valor, se coloca su valor	-1	Reportar los casos al DIEE

	por defecto.		
Empresa			
Nombre Comercial	En caso de no existir un valor, se coloca su valor por defecto.	-1	
Actividad Comercio Exterior	En caso de no existir un valor, se coloca su valor catalogado	Desconocido	Reportar los casos al DIEE
Actividad Económica	Si la actividad económica presenta errores o no tiene una correspondencia directa	No aplica	Reportar los casos al DIEE
Fecha de inscripción	No pasa su validación	No aplica	Reportar los casos al DIEE
Fecha de inicio de actividad	No pasa su validación	No aplica	Reportar los casos al DIEE
Fecha de cese de actividad	En caso de no existir un valor, se coloca su valor por defecto.	-1	
Fecha de reinicio de actividad	En caso de no existir un valor, se coloca su valor por defecto.	-1	
Fecha de actualización	En caso de no existir un valor, se coloca su valor por defecto.	-1	
Unidad Local			
Numero Ruc	No presenta mínimo 13 dígitos	No aplica	El registro no sube y reportar al DIEE
Número de unidad local	Existen unidades locales con el mismo número	No aplica	Reportar los casos al DIEE
Nombre Unidad local	En caso de no existir un valor, se coloca su valor por defecto.	-1	
Actividad Económica	Si la actividad económica presenta errores o no tiene una correspondencia directa	No aplica	El registro se debe reportar al DIEE
Actividad Económica Secundaria	En caso de no existir un valor, se coloca su valor por defecto.	-1	
Producto Elaborado	En caso de no existir un valor, se coloca su valor por defecto.	-1	

Producto Comercializado	En caso de no existir un valor, se coloca su valor por defecto.	-1	
Producto Ofertado	En caso de no existir un valor, se coloca su valor por defecto.	-1	
Materia Prima	En caso de no existir un valor, se coloca su valor por defecto.	-1	
Fecha de cierre	En caso de no existir un valor, se coloca su valor por defecto.	-1	
Fecha de apertura	No pasa su validación	No aplica	Reportar los casos al DIEE
Fecha de inscripción	No pasa su validación	No aplica	Reportar los casos al DIEE
Fecha de actualización	En caso de no existir un valor, se coloca su valor por defecto.	-1	
Fecha de reinicio de actividad	En caso de no existir un valor, se coloca su valor por defecto.	-1	
Unidad local estado	En los casos de que difieran los estados de la unidad local, con la unidad legal	No Aplica	Reportar los casos al DIEE
Unidad local tipo	En caso de no existir un valor, se coloca su valor catalogado	Desconocido	Reportar los casos al DIEE

Valores por defecto para empresas del Ministerio de Educación

Variable	Valor por defecto	Descripción
Clase contribuyente	NA	No aplica
Obligado a llevar contabilidad	NA	No aplica
Actividad comercio exterior	5	No aplica
Forma institucional	7	Institución Pública

Conteos

Los conteos se los debe realizar después de todas las validaciones y limpiezas de datos para poder saber qué cantidad de datos han subido con éxito y así poder ver si existe coherencia en la cantidad de información entre tablas de la base de datos.

Conteos a realizar	
CONTEOS	**VERIFICACIÓN**
Contar el número de unidades legales	El número de unidades legales debe ser igual al número de empresas
Contar el número de empresas	
Contar el número de unidades locales	
Contar el número de unidades locales que son matriz	El número de unidades locales matriz debe ser igual al número de empresas y de unidades legales.
Contar las empresas por estado	Reporte
Contar las unidades legales que tienen el campo de geografía = 0 o = null	Resultado debe ser 0
Contar las unidades locales que tienen el campo de geografía = 0 o = null	Resultado debe ser 0
Empleados	El valor de empleados H , M y Total de empresa debe ser igual a la sumatoria de H, M y Total de unidades locales.
Ventas	El valor de ventas es la sumatoria de ventas tarifa 0%, ventas tarifa 12% y exportaciones netas.
Contar las empresas que tienen el campo de numero_unidades_locales = 0 o = null	Resultado debe ser 0

Contar las empresas que no tienen actividad_economica	Resultado debe ser 0
Contar las empresas que tienen el campo obligado_llevar_contabilidad vacio o null	Resultado debe ser 0
Todas las unidades locales deben contar con número de ruc y número de establecimiento	
Todas las unidades legales y empresas deben contar con número de ruc	

Anexos

En los anexos vamos a poder encontrar documentos que contienen información de cómo se procedió con el trabajo de limpieza de datos, validación de campos e información referente y complementaria a este documento, para la obtención de una información de calidad, es por eso que es importante revisar los anexos y a partir de ellos trabajar con las bases de datos.

Validación del RUC:

Ver Anexo 1 (Validación del RUC)

Tipos de validaciones a realizar:

Ver Anexo 2 (Plan de validaciones automáticas y manuales)

Informe Limpieza de Datos

Ver Anexo 3 (Informes Limpieza de Datos)

Plan de Inconsistencias con cruce de Variables

Ver Anexo 4 (Plan de Inconsistencias con cruce de Variables)

ANEXO 1

Validación automática del RUC

Directorio de Empresas y Establecimientos

VALIDACIÓN DEL RUC

El documento actual trae las validaciones para el ruc tanto de personas naturales como para sociedades.

RUC jurídicos y extranjeros sin Cedula:

- Todas las empresas deben tener un RUC.
- El largo es de 13 números
- Para validar que el ruc sea el correcto se procede de la siguiente manera:

Ejemplo:

Coeficiente: 4 3 2 7 6 5 4 3 2

RUC: 1 7 9 0 0 1 1 6 7 4 0 0 1

Producto: 4 21 18 0 0 5 4 18 14

Sumatoria: 84

Residuo: 84 dividido para 11 da como respuesta 7, y como residuo 7.

Resta: 11 – 7 = 4.

4 Es el dígito verificador.

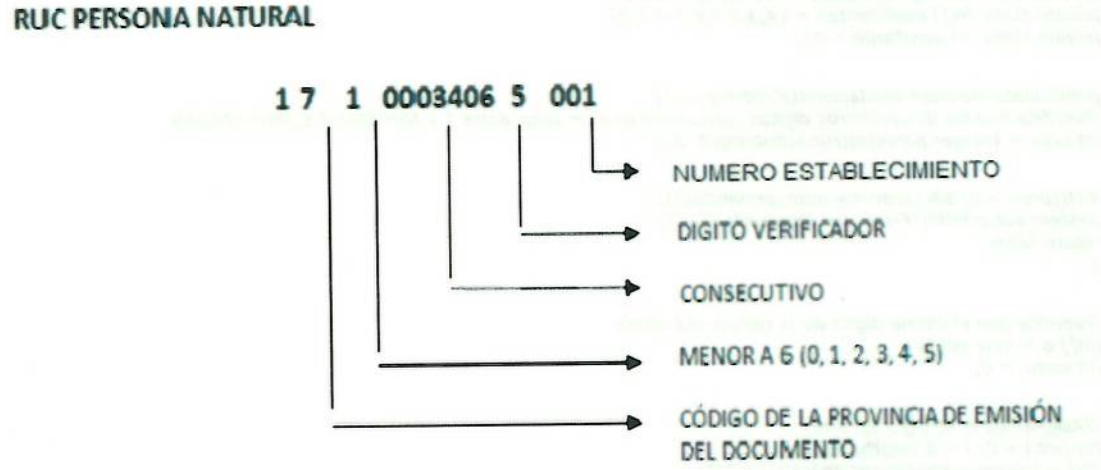
Código de validación en java:

```
public class ValidaRucSociedades {

/**
* @param args
*/
private static  final int num_provincias = 24;
//public static String rucPrueba = "1790011674001";
private static int[] coeficientes = {4,3,2,7,6,5,4,3,2};
private static int constante = 11;

public static Boolean validacionRUC(String ruc){
//verifica que los dos primeros dígitos correspondan a un valor entre 1 y NUMERO_DE_PROVINCIAS
int prov = Integer.parseInt(ruc.substring(0, 2));

if (!((prov > 0) && (prov <= num_provincias))) {
System.out.println("Error: ruc ingresada mal");
return false;
}

//verifica que el último dígito de la cédula sea válido
int[] d = new int[10];
int suma = 0;

//Asignamos el string a un array
for (int i = 0; i < d.length; i++) {
d[i] = Integer.parseInt(ruc.charAt(i) + "");
}

for (int i=0; i< d.length - 1; i++) {
d[i] = d[i] * coeficientes[i];
suma += d[i];
//System.out.println("Vector d en " + i + " es " + d[i]);
}

System.out.println("Suma es: " + suma);

int aux, resp;

aux = suma % constante;
resp = constante - aux;

resp = (resp == 10) ? 0 : resp;

System.out.println("Aux: " + aux);
System.out.println("Resp " + resp);
System.out.println("d[9] " + d[9]);

if (resp == d[9]) {
return true;
}
else
return false;
}

public static void main(String[] args) {
String ruc_dato = "1790011674001";
if (validacionRUC(ruc_dato))
System.out.println("El RUC es correcto");
else
System.out.println("El RUC es incorrecto");
}
}
```

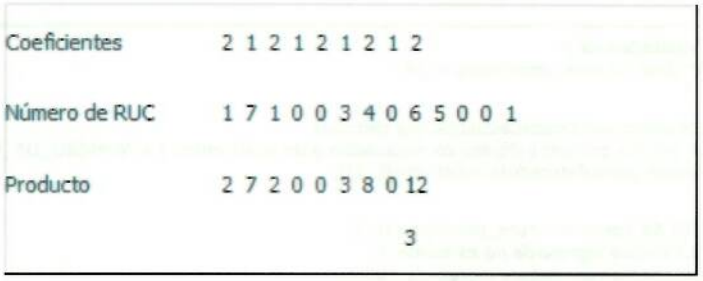
- RUC de personas naturales

RUC persona natural:

La estructura de este tipo de RUC se muestra en la siguiente figura:



- El RUC de una persona natural será 13 dígitos, sin letras, sin caracteres especiales, únicamente números, de los cuales los 10 primeros será la cédula de identidad.
- Las 2 primeras posiciones corresponden a la provincia donde fue expedida, por lo cual los dos primeros números no será mayor a 24 ni menor a 1.
- El tercer dígito es menor a 6 (0,1,2,3,4,5).
- Del cuarto al noveno dígito es un número consecutivo de 6 dígitos.
- El décimo dígito es el dígito verificador.
- Del décimo primer dígito al décimo tercer dígito, identifican en forma consecutiva el número de establecimientos. Empieza siempre con el 001.



Como se ve en la imagen se usan los 9 primeros dígitos del RUC. Los números de las posiciones impares e multiplican por 2, y los números de las posiciones pares se multiplican por 1. Si el resultado de multiplicar por 2 es mayor a 9, se resta 9. Como en la figura, 12 – 9 = 3.

Luego se suman todos los valores del producto. En este caso la respuesta es 25.

Luego se resta este número de su decena superior. La decena superior a 25 es 30. Por lo tanto queda: 30 – 25 = 5.

Y 5 es el dígito verificador, si este número coincide con el décimo dígito del RUC la cédula es correcta.

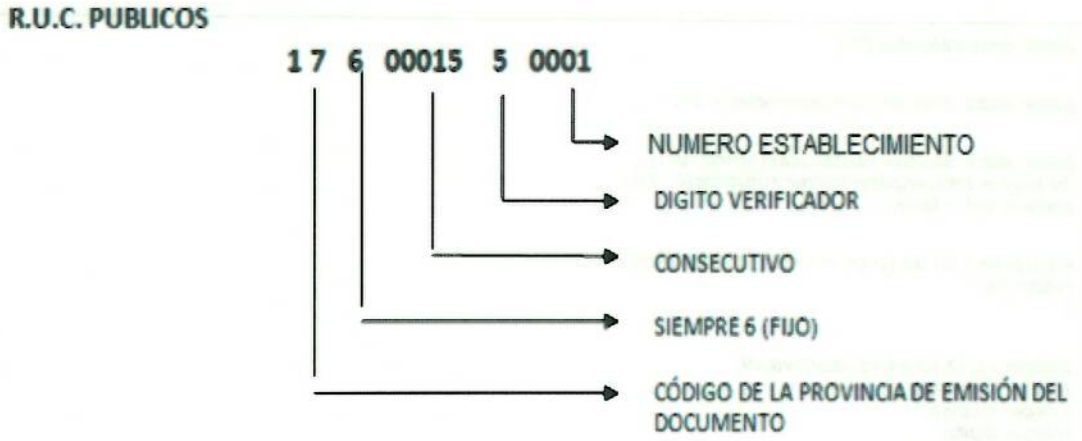
Excepción: Si el resultado de la resta es 10, el dígito verificador será 0.

Usando el código de la página que les puse arriba y haciéndole una pequeña modificación, el resultado queda:

```
public class ValidaCedula {
private static  final int num_provincias = 24;


public static Boolean validacionCedula(String cedula){
//verifica que los dos primeros dígitos correspondan a un valor entre 1 y NUMERO_DE_PROVINCIAS
int prov = Integer.parseInt(cedula.substring(0, 2));


if (!((prov > 0) && (prov <= num_provincias))) {
//addError("La cédula ingresada no es válida");
System.out.println("Error: cedula ingresada mal");
return false;
}

//verifica que el último dígito de la cédula sea válido
int[] d = new int[10];
//Asignamos el string a un array
for (int i = 0; i < d.length; i++) {
d[i] = Integer.parseInt(cedula.charAt(i) + "");
}

int imp = 0;
int par = 0;

//sumamos los duplos de posición impar
for (int i = 0; i < d.length; i += 2) {
d[i] = ((d[i] * 2) > 9) ? ((d[i] * 2) - 9) : (d[i] * 2);
imp += d[i];
}


//sumamos los digitos de posición par
for (int i = 1; i < (d.length - 1); i += 2) {
par += d[i];
}

//Sumamos los dos resultados
int suma = imp + par;


//Restamos de la decena superior
int d10 = Integer.parseInt(String.valueOf(suma + 10).substring(0, 1) +
"0") - suma;


//Si es diez el décimo dígito es cero
d10 = (d10 == 10) ? 0 : d10;


//si el décimo dígito calculado es igual al digitado la cédula es correcta
if (d10 == d[9]) {
return true;
}else {
//addError("La cédula ingresada no es válida");
return false;
}
}


public static void main(String[] args) {
String ruc_dato = "1710034065001";
if (validacionCedula(ruc_dato.substring(0, 10)))
System.out.println("El RUC es correcto");
else
System.out.println("El RUC es incorrecto");
}
}
```

RUC para empresas públicas:

La estructura de este tipo de RUC se muestra en la siguiente figura:



La validación de este tipo de RUC se basa en el algoritmo de Módulo 11. Los coeficientes son: 3, 2, 7, 6, 5, 4, 3, 2.

El procedimiento es el siguiente:

- ✓ Se multiplican los primeros nueve dígitos del RUC por cada uno de los coeficientes que le corresponde en la misma posición.
- ✓ Se suma ese resultado.
- ✓ Se divide ese resultado para el módulo, como este RUC es módulo 11, se divide la suma para 11, y se obtiene su residuo.
- ✓ Se resta el módulo (en este caso 11) de el residuo de la división anterior.
- ✓ El resultado es el dígito verificador. Si ese número coincide con el número del RUC de la posición 9 el RUC es correcto.

Excepción: Si el residuo es cero, el dígito verificador es cero.

Ejemplo:

Coeficiente:	3	2	7	6	5	4	3	2					
RUC:	1	7	6	0	0	0	1	5	5	0	0	0	1
Producto:	3	14	42	0	0	0	3	10	14				
Sumatoria:	72												

Residuo: 72 dividido para 11 da como respuesta 6, y como residuo 6.

Resta: 11 – 6 = 5.

5 Es el dígito verificador.

El código es el siguiente:

```
public class ValidaRucEP {

public static  final int num_provincias = 24;

public static Boolean validaRucEP(String ruc){
int prov = Integer.parseInt(ruc.substring(0, 2));
boolean val = false;

if (!((prov > 0) && (prov <= num_provincias))) {
return val;
}

Integer v1,v2,v3,v4,v5,v6,v7,v8,v9;
Integer sumatoria;
Integer modulo;
Integer digito;
Integer sustraendo;
int[] d = new int[ruc.length()];

for (int i = 0; i < d.length; i++) {
d[i] = Integer.parseInt(ruc.charAt(i) + "");
}

v1 = d[0]* 3;
v2 = d[1]* 2;
v3 = d[2]* 7;
v4 = d[3]* 6;
v5 = d[4]* 5;
v6 = d[5]* 4;
v7 = d[6]* 3;
v8 = d[7]* 2;
v9 = d[8];

sumatoria = v1+v2+v3+v4+v5+v6+v7+v8;
modulo = sumatoria % 11;
sustraendo = modulo * 11;
digito = 11-(sumatoria – sustraendo);
System.out.println("Digito RUC      -> "+digito);
System.out.println("Digito Calculado -> "+v9);

if(digito == v9){
val = true;
}else
val = false;
return val;
}

public static void main(String[] args) {
String ruc_dato = "1760001550001";
if (validaRucEP(ruc_dato)) {
System.out.println("El RUC es correcto");
} else
System.out.println("El RUC es incorrecto");
}
}
```

ANEXO 2

Plan de Validaciones Automáticas y Manuales

Directorio de Empresas

V 5.1.0.0

Contenido

Introducción

Desarrollar un plan de validación el cual sirva como guía de validaciones para el Directorio de Empresas y Establecimientos DIEE, tiene como principal objetivo dejar en claro validaciones que son necesarias realizar para asegurar que la información con la que se trabaja sea de calidad. Estas validaciones son las que con más frecuencia se han venido realizado a lo largo del tiempo.

Existen validaciones que se las ha manejado mediante Excel, scripts a la base de datos, integrando en el proceso ETL y recientemente la herramienta SQL PowerDQguru. Para lo cual se ha necesitado la participación del equipo de Call Center, analistas económicos y técnicos informáticos.

En el presente documento se priorizará el aspecto de la integración de todas estas validaciones en un proceso automático que pueda englobar la mayor parte de ellas.

Análisis de herramientas de software para validación y edición de datos.

Debido a la gran cantidad de datos con el que trabaja el Directorio de Empresas y Establecimientos es necesario contar con una o varias herramientas de apoyo en el proceso de validación, para lo cual se debe realizar una investigación de las herramientas de software libre que permitan realizar este trabajo de forma automática.

Pentaho

Pentaho es una herramienta de Business Intelligence desarrollada bajo la filosofía del software libre para la gestión y toma de decisiones empresariales. Es una plataforma compuesta de diferentes programas que satisfacen los requisitos de BI. Ofreciendo soluciones para la gestión y análisis de la información, incluyendo el análisis multidimensional OLAP, presentación de informes, minería de datos y creación de cuadros de mando para el usuario.

La plataforma ha sido desarrollada bajo el lenguaje de programación Java y tiene un ambiente de implementación también basado en Java, haciendo así que Pentaho sea una solución muy flexible al cubrir una alta gama de necesidades empresariales.

Los productos destacados ofertados en la Suite de Business Intelligence son los siguientes:

- Pentaho Data Integration: Herramienta que proporciona mediante una interfaz de usuario sencilla e intuitiva la posibilidad de manipulación de los datos desde una fuente externa e independiente a la herramienta.
- Pentaho Analisys Services: Herramienta para crear cubos multidimensionales OLAP. Soporta el lenguaje de consulta MDX (expresiones multidimensionales) y lenguaje XML para el análisis y especificaciones.
- Pentaho Reporting: Herramienta con la cual el usuario será capaz de crear informes usando datos de fuentes externas. Estos informes son generados en XML y pueden ser exportados a diversos tipo de archivos finales, como puede ser PDF, HTML o documentos de texto. Una de las características es que dispone de un menú interactivo que guía al usuario paso por paso en la creación de los informes.
- Pentaho Data Mining: Herramienta para extraer información implícita en los datos. Desarrollado con el motor de minería de datos Weka. Permite extraer patrones, clusterizar, clasificar o extraer reglas de asociación de los datos.
- Pentaho DashBoard: Herramienta para crear cuadros de mando en la interfaz final de la herramienta web. Estos cuadros de mando podrán realizar funciones de consulta y análisis de los datos.
- Pentaho BI Server: Herramienta que proporciona el servidor y plataforma web del usuario final. Este podrá interactuar con la solución Business intelligence previamente creada con las herramientas anteriormente comentadas.

Variables involucradas

Existe una gran cantidad de variables dentro del Directorio, de las cuales se destacarán aquellas más relevantes.

Matriz de prioridad

Prioridad	Descripción
1	Muy Alta
2	Alta
3	Media
4	Baja

TABLA 1TABLA DE PRIORIDADES

Variable (campo a validar)	Esquema: Diemp Tabla	Validaciones	Prioridad
fecha_inicio_actividades	f_empresa	Validación de que exista un valor. Caso contrario se coloca (-1). Validación de que sean 8 números consecutivos. Validación de que sea aaaammdd.	3
fecha_cese_actividad	f_empresa	Validación de que exista un valor. Caso contrario se coloca (-1). Validación de que sean 8 números consecutivos. Validación de que sea aaaammdd.	3
id_actividad_comercio_exterior	f_empresa	Por implementar	3
nombre_comercial	f_empresa	Validación de que exista un valor. Caso contrario se coloca (-1). Valida que por lo menos exista 3 caracteres.	3
fecha_reinicio_actividad	f_empresa	Validación de que exista un valor. Caso contrario se coloca (-1). Validación de que sean 8 números consecutivos. Validación de que sea aaaammdd.	3
fecha_inscripcion	f_empresa	Validación de que exista un valor. Caso contrario se coloca (-1). Validación de que sean 8 números consecutivos. Validación de que sea aaaammdd.	3
fecha_actualizacion	f_empresa	Validación de que exista un valor. Caso contrario se coloca (-1). Validación de que sean 8 números consecutivos. Validación de que sea aaaammdd.	3
id_ruc	f_unidad_legal	Validación de número de ruc	1
unidad_legal_estado	f_unidad_legal	Validación de que exista un valor.	1
unidad_legal_tipo	f_unidad_legal	Validación de que exista un valor.	1

expediente	f_unidad_legal	Validación de que exista un valor. Caso contrario se coloca (-1).	4
id_acto_juridico	f_unidad_legal	Validación de que exista un valor. Caso contrario se coloca un valor por defecto.	4
ruc_adscrita	f_unidad_legal	Validación de que exista un valor. Caso contrario se coloca (-1).	4
id_clase_contribuyente	f_unidad_legal	Validación de que exista un valor. Caso contrario se coloca un valor por defecto.	1
fecha_acto_juridico	f_unidad_legal	Validación de que exista un valor. Caso contrario se coloca (-1). Validación de que sean 8 números consecutivos. Validación de que sea aaaammdd.	2
ruc_acto_juridico	f_unidad_legal	Validación de que exista un valor. Caso contrario se coloca (-1).	2
obligado_llevar_contabilidad	f_unidad_legal	Validación de que exista un valor. Caso contrario se coloca (-1).	1
id_clasificacion_fjuridica	f_unidad_legal	Validación de que exista un valor. Caso contrario se coloca un valor por defecto.	2
id_forma_juridica	f_unidad_legal	Validación de que exista un valor. Caso contrario se coloca un valor por defecto.	2
cod_ubicacion_sri	f_unidad_local	Validación de que exista un valor y tenga dato hasta parroquia.	1
id_ruc	f_unidad_local	Validación del número de ruc.	1
numero_unidad_local	f_unidad_local	Valida que no exista el número cero.	1
unidad_local_nombre	f_unidad_local	Validación de que exista un valor. Caso contrario se coloca (-1).	1
fecha_cierre	f_unidad_local	Validación de que exista un valor. Caso contrario se coloca (-1). Validación de que sean 8 números consecutivos. Validación de que sea aaaammdd.	3
fecha_desde	f_unidad_local	Validación de que exista un valor. Caso contrario se coloca (-1). Validación de que sean 8 números consecutivos. Validación de que sea aaaammdd.	3
fecha _apertura	f_unidad_local	Validación de que exista un valor. Caso contrario se coloca (-1). Validación de que sean 8 números consecutivos. Validación de que sea aaaammdd.	3
Id_actividad_economica	f_unidad_local	Validación de que exista un valor.	1
id_unidad_local_estado	f_unidad_local	Validación de que exista un valor.	1
unidad_local_tipo	f_unidad_local	Validación de que exista un valor y que exista una sola unidad local matriz por empresa.	1

cod_ubicacion_sri	f_unidad_local	Validación de que exista un valor y tenga dato hasta parroquia.	1
fecha_reinicio_actividad	f_unidad_local	Validación de que exista un valor. Caso contrario se coloca (-1). Validación de que sean 8 números consecutivos. Validación de que sea aaaammdd.	3
fecha_inscripcion	f_unidad_local	Validación de que exista un valor. Caso contrario se coloca (-1). Validación de que sean 8 números consecutivos. Validación de que sea aaaammdd.	3
fecha_actualizacion	f_unidad_local	Validación de que exista un valor. Caso contrario se coloca (-1). Validación de que sean 8 números consecutivos. Validación de que sea aaaammdd.	3
descripcion	f_medio_comunicacion	Validaciones realizadas a medios de comunicación.	2
descripcion	f_ubicacion_direccion	Limpieza, validación y estandarización realizadas a direcciones	2

Se deberá plantear una solución para todas aquellas variables cuya prioridad sea 1 y 2. Se planteará en lo posible una solución para todas aquellas variables que su prioridad sea 3 y 4.

Tipos de validaciones a realizar

Es necesario clasificar las validaciones ya que se optimizará la forma de validar a las variables.

A continuación se describe las validaciones que se puede realizar y se colocará un código el cual se deberá mantener a lo largo del proceso.

<table>
<tr><th>Código</th><th>Validación</th><th colspan="3">Descripción</th><th>Complejidad</th></tr>
<tr><td>1</td><td>Números</td><td colspan="3">Valida que el campo solo contenga números.</td><td>Baja</td></tr>
<tr><td>2</td><td>Longitud</td><td colspan="3">Valida que exista una longitud mínima de caracteres.</td><td>Baja</td></tr>
<tr><td>3</td><td>Null</td><td colspan="3">Valida que no tenga valores null.</td><td>Baja</td></tr>
<tr><td>4</td><td>Vacios</td><td colspan="3">Valida que el campo no contenga espacios vacios.</td><td>Baja</td></tr>
<tr><td>5</td><td>Fechas</td><td colspan="3">Valida que tenga 8 dígitos.</td><td>Media</td></tr>
<tr><td>6</td><td>RUC</td><td colspan="3">Valida que sea un Ruc.
1ro y 2do dígito | 17 | CÓDIGO DE LA PROVINCIA DE EMISIÓN DEL DOCUMEN
3er Dígito | 9 | SIEMPRE 9 cuanto son Personas Jurídicas y extranjeros cédula.
3er Dígito | 6 | SIEMPRE 6 cuanto son entes Públicos.
3er Dígito | (0-5) | Cuando son Personas Naturales.</td><td>Alta</td></tr>
</table>

		Del 4to al 9no dígito	001167	CONSECUTIVO	
		10mo dígito	4	DIGITO VERIFICADOR	
		Del 11ro al 13ro dígito	001	RUC	
7	Si / No	Valida que el campo contenga 1 solo carácter y sea S o N.			Alta
8	Excluyente	Valida que el campo no contenga cierta información.			Media
Geografía					
9	Geografía	Valida que exista información hasta parroquia.			Alta
10	Geografía y zonificación	Valida que la codificación de geografía y zonificación existan dentro de su catálogo respectivo.			Alta
Actividad Económica					
11	CIIU4	Valida que la actividad debe corresponder a su código.			Alta
Empleados					
12	Sumatoria empleados 1	Valida que la suma del personal afiliado de una empresa debe ser igual a la suma del personal hombre más mujer.			Alta
13	Sumatoria empleados 1	Valida que la suma de los empleados de las unidades locales sea igual a la cantidad de empleados que reporta la empresa.			Alta
Remuneraciones					
14	Sumatoria remuneraciones	Valida que la suma de remuneraciones de una empresa debe ser igual a la suma de remuneraciones de hombres más mujeres.			Alta
Unidad local					
15	Unidad local	Valida que una empresa por lo menos debe tener 1 unidad local.			Alta
16	Matriz	Valida que el número de unidades locales matrices, sea igual al número de empresas.			Alta

Validaciones Automáticas

Las validaciones automáticas son aquellas que mediante el uso de una herramienta de software se las podrá solventar.

Existen dos tipos de validaciones dentro de este grupo.

- Validaciones dentro del proceso ETL.
- Validaciones fuera del proceso ETL.

Validaciones Automáticas dentro del proceso ETL

Estas validaciones son aquellas que se realiza por cada uno de los registros.

- Validación de longitud del campo.
- Validación de campos con valores null y vacíos.
- Validación de números.
- Validación de caracteres.
- Validación de formato de fechas.

Validaciones Automáticas fuera del proceso ETL

Estas validaciones se deberán ejecutar una vez finalizado el proceso ETL. Entre estas se encuentran comúnmente conteos de totales de empresas, totales de unidades legales y unidades locales, entre otras que se detallarán a continuación.

- Validación de que coincida el número de unidades locales matrices con el número de empresas y el número de unidades legales.
- Validación de que coincida el número de empleados en la empresa con la suma de los empleados de sus unidades locales.

Otras Validaciones Automáticas

Existen otras validaciones que se relacionan con limpieza de datos, se debe tomar en cuenta que este tipo de validaciones no se las puede incluir dentro del proceso ETL, debido a que el proceso de naturaleza es pesado al ejecutarse y al incrementar validaciones de cada uno de los registros en los que aparte de validar se deba modificar la información, causará que el proceso colapse.

Es importante notar que se deben realizar la limpieza de datos una vez que los datos se encuentren cargados, con el fin de completar el ciclo de calidad en la información del directorio, este tipo de validaciones se dan en los siguientes campos y por los aspectos a continuación descritos.

Teléfono:

- Validación del número de dígitos — Se puede Automatizar para que valide que el campo contenga mínimo 6 números consecutivos.
- Si el campo tiene 6 dígitos — Se debe completar la información.
- Si el campo tiene 7 dígitos — Se debe completar con el código provincial.
- No se puede validar que solo sean números debido a que en el campo también existen números de teléfono que tienen extensiones. Por ejemplo: 2611464 ext.123

Este tipo de validaciones en donde se debe modificar la información no se la puede realizar en el proceso ETL. Ya que este solo valida si según lo especificado deja pasar la información o no. No la modifica debido a que incluir en el proceso ETL modificación de datos, cargaría demasiado al proceso que ya de por si es largo y ocupa altos recursos del servidor.

Validaciones Manuales

Las validaciones manuales son las que se encarga de realizar el equipo de Análisis o Call Center, que son a quienes se entrega un listado de empresas que se han encontrado algún tipo de inconsistencia y es necesario una revisión uno a uno de los casos encontrados, esta validación es retroalimentada a la BDD del DIEE en la etapa de procesamiento.

Otras Validaciones Manuales

Otras validaciones que por ser demasiado inconstantes se vuelve un proceso casi manual, es la limpieza, separación y estandarización de direcciones, donde existe una cantidad infinita de posibles abreviaciones para los diferentes detalles de las mismas. Otro de los principales inconvenientes con direcciones es cuando alguna de las fuentes nos entrega la dirección en un solo campo y toca buscar mecanismos y parámetros de separación en los 12 campos con los que trabaja el DIEE.

Conclusiones

- Existen varias herramientas en el mercado, las cuales pueden satisfacer en gran parte las necesidades actuales de los procesos de validaciones de información del DIEE, cada una con sus respectivas ventajas y desventajas. La decisión de optar por una herramienta toma en cuenta ciertos factores como son la usabilidad, tiempo de aprendizaje, costos, soporte al usuario, actualizaciones, capacitación, integración con otras herramientas, entre otras que deben ser evaluadas en conjunto.
- Se debe tener presente que existen herramientas propietarias las cuales pueden aportar en gran manera para satisfacer las necesidades del DIEE, mejorando en aspectos como tiempo de desarrollo, facilidad de interacción entre usuario y aplicación, mejoras en dashboards, posibilidad de crear informes tipo infografías como los que se necesita para las publicaciones, entre otros aspectos, características que en herramientas de software libre no se pude encontrar, o se tiene en menor calidad.
- Se debe tener en cuenta que existen varios factores para poder implementar las validaciones descritas, entre ellos son las herramientas que se utilice, el tiempo de capacitación de las mismas, curva de aprendizaje, la complejidad de las validaciones y el número de campos que se va a validar.

Recomendaciones

- No solo la investigación es suficiente para poder manejar alguna herramienta, porque al auto-capacitarse no se puede explotar al máximo las herramientas, es por eso que se recomienda tener capacitación en cuanto las herramientas.
- Es importante tomar en cuenta que en el camino para generar una información de calidad, se debe pasar por varios procesos uno de los cuales es el hacer validaciones, el cual no garantiza que se obtenga un 100% de información confiable, pero si una notable mejora.
- Se debe considerar que es indispensable la participación del personal de análisis de información, en el proceso de validación de datos, esto debido a que existe la posibilidad de que cambien las reglas del negocio y se necesite que se tomen nuevas decisiones las cuales se vean reflejadas en los procesos de validación.

ANEXO 3

Limpieza de Datos

Directorio de Empresas y Establecimientos

INTRODUCCIÓN

Para el proceso de limpieza de la información que se ha optado por realizarlo previo a la etapa de procesamiento de la información con el fin de tener información netamente significante y de mejor calidad para ser adherida a la Base de Datos del DIEE.

Para obtener mejores resultados en el proceso de limpieza se ha optado por la utilización de la herramienta SQL Power DQguru, funcionalidad que se detalla en el siguiente apartado.

SQL Power DQguru

El SQL Power DQGuru es una herramienta de Limpieza de Datos que, el fabricante SQLPower ha liberado convirtiendo la licencia en Open Source. Pero esta herramienta no solo realiza una limpieza de datos, sino que también valida y corrige direcciones, identifica y elimina duplicados y crea referencias cruzadas entre las tablas de origen y destino. Esto proporciona a los usuarios de negocios datos completos y precisos, y una visión completa de todas las entidades que son analizadas mediante este producto.

Esta Herramienta tiene las siguientes características:

- Interfaz gráfica de usuario intuitiva que permite la adopción y el uso rápido para los analistas de datos.
- Interfaz de proceso intuitivo denominado 'transformación' que permite crear e implementar rápidamente los flujos de trabajo de conversión de datos.
- Los usuarios pueden definir sus propios datos si coinciden con los criterios.
- Puede ser utilizado para los datos iniciales o periódicos de limpieza.
- Genera tablas de referencias cruzadas para vincular los identificadores del sistema de origen a los identificadores de la base de datos de destino.
- Amplio soporte para funciones de transformación y coincidentes:
    - Concatenación, que permite la unión de cadenas de texto de diferentes campos.

- Doble Metaphone, Metaphone, refinados soundex, codificación fonética soundex, que son objetos para comparación de cadenas de texto, mediante la utilización de algoritmos fonéticos.
- Conversión de mayúsculas o minúsculas del alfabeto.
- Sustitución de cadenas.
- Subcadena y subcadena de palabra.
- Sustitución de palabras a través de la traducción de las palabras o grupos de palabras.

- Distintos niveles de transformación de datos también son compatibles para ayudar a administrar el desarrollo y ejecución de los procesos de limpieza de datos:
  - El motor de coincidencia identifica duplicados, almacenando los resultados en una tabla de destino, sin modificar los datos de origen.
  - El motor de fusión elimina los registros duplicados de los datos de origen de acuerdo con las reglas que ha definido.
  - El motor de limpieza sustituye a los registros de los datos de origen con datos reformateados de acuerdo con sus normas.
- Amplio soporte para varias bases de datos, para los datos de origen y destino.

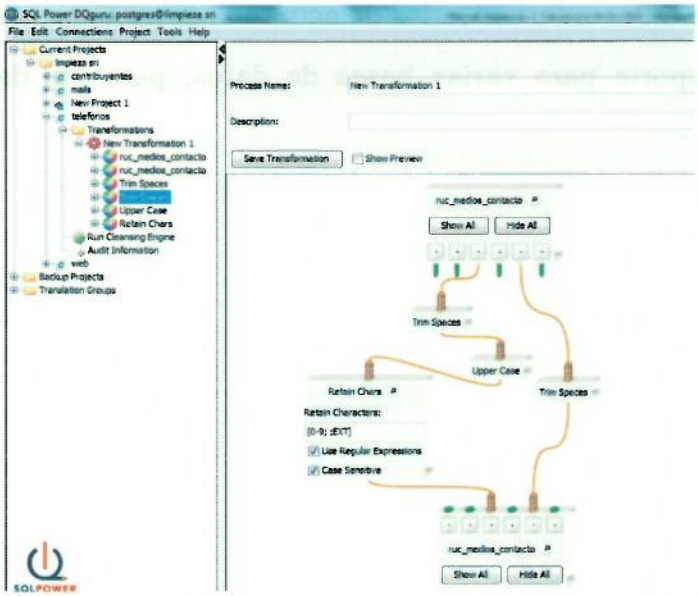
PROCESO DE LIMPIEZA DE VARIABLES

Una vez obtenida la información de las diferentes fuentes, el procedimiento para la limpieza de datos a seguir es el siguiente:

- Congelamiento de la base nueva
- Partir de esta nueva base con la respectiva limpieza, y validación de datos.
- Cargar nueva información a la base del DIEE.

Para el proceso de limpieza de la información se utiliza la herramienta SQL Power DQguru, con la cual se eliminan todos los caracteres extraños o no permitidos en un campo determinado, es importante mencionar que esta herramienta, a diferencia de otras, no solamente realiza un análisis de la información que necesita ser evaluada sino que ataca directamente a la misma eliminando la información innecesaria o errónea. Este proceso se lo realiza especialmente a la tabla de medios de comunicación donde se tiene gran cantidad de caracteres innecesarios que no son permitidos por ejemplo en los números telefónicos. En el resto de tablas también se realiza una limpieza para eliminar los espacios en blanco principalmente.

En el siguiente gráfico se muestra un ejemplo:



A continuación se destacan los procedimientos que se debe realizar a toda la información que se la va a dar una limpieza de datos:

- Colocar los datos en sus respectivas catalogaciones:

- Todos los datos que contengan las palabras www. ó http. Se colocará en tipo web.
- Todos los datos que contengan el símbolo @. Se colocará en tipo correo electrónico.

- Colocar todo en mayúsculas con excepción de datos pertenecientes a correos y páginas web.

Números telefónicos:

Esta limpieza de datos se lo realizará a todos los campos que contengan información del número de teléfono de la siguiente manera:

- Se debe borrar todo dato que no tenga por lo menos un número y no esté compuesto solo por cero.
- Se debe borrar los espacios en blanco.
- Se debe borrar los paréntesis para abrir "(" y cerrar ")"
- Se debe borrar la palabra "REF"
- Se debe borrar el signo punto ".", slash "/", asterisco "*", guión "-", mas "+", llaves "{" y "}", dos puntos ":"
- Se debe borrar los datos de teléfono cuyo largo sea igual a 9 y no comiencen con: 0 ó 9.
- Se debe borrar toda letra con excepción de datos que contengan la palabra "EXT".
- Se debe borrar los datos de números que terminen en 000000 cuya longitud sea menor a 10.
- Se debe borrar los datos cuya longitud sea igual a 10, no contengan la palabra "EXT" y comiencen con 02, 03, 04, 05, 06 ó 07. Previo a esto revisar si los dígitos 3 y 4 son iguales, en este caso eliminar el dígito 3 para recuperas éstos casos.
- Cambio de tipo de medio de contacto.

  - Se debe cambiar a tipo celular a los datos cuya longitud es 8 y comienzan con 9 u 8.
  - Se debe cambiar a tipo celular a los datos cuya longitud es 9 y comienzan con 9.
  - Se debe cambiar a tipo celular a los datos cuya longitud es 9 y comienzan con 09 o 08.

- Se debe cambiar a tipo celular a los datos cuya longitud es 10 y comienzan con 09.

- Los números que tienen solo 6 dígitos, se guardan con tipo de contacto = 9 y quedan pendientes a ser revisados.
- Se debe aumentar el código provincial a los números que solo tengan 7 dígitos.
- Los números de teléfono cuya longitud es igual a 10 y el tercer y cuarto dígito son iguales, reemplazar esta duplicidad por un solo dígito, para así recuperar posibles números correctos.
- Se debe borrar todo dato cuya longitud sea menor a 6 dígitos.
- Se debe aumentar el cero "0" a los datos tengan 8 dígitos e inicien con 2, 3, 4, 5, 6 o 7.
- Se debe aumentar el número 2 luego del código provincial de teléfono para los números de 8 dígitos que coincidían sus dos primeros dígitos con los dígitos de código provincial de teléfono.
- Se debe borrar los números cuya longitud sea de 6 dígitos y empiece con cero.
- Cuando el número comienza con 593 y la longitud es mayor a 8 se borran los 3 dígitos iniciales.
- Datos que presentan la palabra EXT

  - Se debe borrar los datos en donde la palabra EXT esté en una posición menor a 7.
  - Se debe borrar los datos en donde la palabra EXT esté en una posición mayor a 10.

- Se debe aumentar el código provincial a los números con 7 dígitos.
- Para los datos cuya longitud es igual a 8, comienzan en cero y no concuerdan sus dos primeros dígitos con el código de teléfono provincial. Se debe colocar en el campo estado el valor de cero "0", que signifique que la información está inactiva.
- Se debe borrar los datos de números que terminen en 000000 y empiecen con 09990, 0990 y 0910.
- *NOTA: esta validación también aplica a números telefónicos de FAX*

Celulares:

- Se debe borrar todo dato que no tenga por lo menos un número y no esté compuesto solo por cero.
- Se debe borrar los espacios en blanco.
- Se debe borrar los signos punto ".", slash "/", asterisco "*", guión "-", mas "+", llaves "{" y "}", dos puntos ":"
- Se debe borrar toda letra.
- Se debe borrar los datos que tengan longitud mayor a 10.
- Se debe borrar los datos que tengan longitud menor a 8.
- Se debe aumentar el "09" a los datos de 8 dígitos que comiencen en 9 u 8.
- Se debe aumentar el "09" a los datos de 9 dígitos que comiencen en 0.
- Se debe aumentar cero a todos los datos de 9 dígitos que no comiencen en cero.
- Se debe borrar todos los datos de 10 dígitos que no comiencen en cero.
- Para los datos cuya longitud es igual a 8, se deber colocar en el campo estado el valor de cero "0", que signifique que la información está inactiva.

Correo Electrónico:

Esta limpieza de datos se la debe realizar de la siguiente manera:

Se debe borrar todo dato que por lo menos no contenga el símbolo "@"

- Se debe borrar todo espacio en blanco.
- Se debe borrar todo dato cuya longitud sea menor a 6.
- Se debe borrar todo dato que no contenga por lo menos una letra.

Web:

Esta limpieza de datos se la debe realizar de la siguiente manera:

Se debe borrar todo espacio en blanco.

- Se debe borrar todo dato cuya longitud sea menor a 3.
- Se debe borrar todo dato que no contenga por lo menos una letra.

Nombre de Contacto:

Esta limpieza de datos se la debe realizar en los datos de los nombres de contactos de las empresas de la siguiente manera:

- Se debe borrar todo número.
- Se debe verificar que se tenga una longitud mínima de 3.
- Se debe borrar los espacios en blanco al principio y al final de los datos.

Detalle geografía:

Se debe validar los campos: provincia, cantón, parroquia y nombre.

- Para la provincia se validará: Que exista por lo menos 3 caracteres y contenga letras de la A-Z se incluye la letra Ñ.
- En caracteres especiales se permite los siguientes: (. \ / Á É Í Ó Ú , - ( ))
- Para cantón parroquia y nombre se admite los anteriores y también números.
- Se debe revisar la base de datos y verificar que no existan símbolos extraños estos campos (como por ejemplo BA□OS DE AGUA SANTA).

Razón social y Nombre comercial:

- Se verificará con el SRI que la razón social y nombre comercial estén correctamente escritos.
- Se debe verificar que se tenga una longitud mínima de 3.
- Además se debe permitir los caracteres que contenga Ñ, tildes, comillas, @, &, puntos, números, signo de suma +, \, /, #, À,È,Ì,Ò,Ù,Ä,Ë,Ï,Ö,Ü , !, _ , °.
- Reemplazar los caracteres extraños correspondientes a problemas de la letra Ñ (Ã‘), letra Í (Ã), apostrofe '(Â´).

NOTA: Para todos aquellos casos que no se engloban en los mencionados se procedió a borrar esa información y dejar el campo vacío.

Limpieza de Direcciones

El proceso de limpieza de direcciones, se lo realiza empleando script's en PostgreSQL, debido a que las variables correspondientes a direcciones poseen características cambiantes, lo que vuelve más complicado la automatización de la limpieza para estos casos. Por lo cual se han establecido una serie de reglas principales como base para obtener una mejor calidad de dicha información.

1. *Resoluciones sobre Direcciones*

- Cuando se tiene datos en: calle_principal, interseccion,carretero,camino se tiene los siguientes casos:
  - CASO 1: cuando se tenga en interseccion '%MARGEN%' se concatena:
    CALLE_PRINCIPAL=CALLE_PRINCIPAL||INTERSECCION,
    INTERSECCION=CARRETERO||CAMINO
  - CASO 2: cuando se tenga en interseccion not like '%MARGEN%' se concatena: REFERENCIA=REFERENCIA||CARRETERO||CAMINO
- Cuando no se tiene datos en calle_principal se concatena CALLE_PRINCIPAL=CARRETERO||CAMINO
- Cuando no se tiene datos en interseccion pero si en calle_principal, se concatena: INTERSECCION=CARRETERO||CAMINO
- Cuando no se tiene datos en calle_principal ni en interseccion se concatena CALLE_PRINCIPAL=CARRETERO||CAMINO
- Las abreviaturas serán trabajadas como limpieza para las direcciones.
- Si se tiene en camino like 'VIA%' y calle_principal is null and intersección is null, pasar respectivamente carretero, camino a calle_principal, interseccion.
- Si se tiene calle_principal like '%PRINCIPAL%', carretero is not null y camino like 'VIA%' pasar la información a calle principal.
- Si se tiene en carretero like 'VIA%' y en camino not like 'MARGEN%' y calle_principal, interseccion is null carretero pasa a calle principal y camino a referencia.
- Si se tiene camino like 'SITIO%' o like 'SECTOR%', carretero like 'VIA%' y calle_principal, interseccion is null, pasar carretero a calle principal y camino a sector

2. ***Transformación de Direcciones actuales a 12 campos***

Se construye una temporal a partir de las direcciones que se obtienen de la fuente, concatenando los campos para reducirlos a 12, los campos a concatenar son:

- edificio_bloque
- manzana_supermanzana
- carretero_camino

3. ***Estandarización de direcciones***

En esta parte lo que se estandariza son las abreviaturas que se encuentran en los diferentes detalles de las direcciones, para que tenga un mismo formato, las abreviaciones son las siguientes:

NUM	NOMBRE DE LA PALABRA	ABREVIATRURA	SIN ABREVIATURA
1	ASOCIACIÓN	ASOC.	
2	AUTOPISTA	AUTOP.	
3	AVENIDA	AV.	
4	BLOQUE		X
5	BODEGA		X
6	BOULEVAR	BULV.	
7	CALLE		X
8	CALLEJÓN	CJON.	
9	CARRETERO/A		X
10	CAMINO		X
11	CASA		X
12	CENTRO COMERCIAL	CC.	
13	CIRCUNVALACIÓN	CIRCUNV.	
14	CIUDADELA	CDLA.	
15	COMUNA		X
16	CONDOMINIO		
17	CONJUNTO	CJTO.	
18	COOPERATIVA	COOP.	
19	DEPARTAMENTO	DPTO.	
20	EDIFICIO	EDF.	
21	EJE VIAL	EJE	
22	GALPÓN		X
23	HACIENDA	HCDA.	
24	KILÓMETRO	KM.	
25	LOCAL		X
26	MEZANINE	MZN	
27	MANZANA	MZ.	
28	NUMERO DE LOCAL	NUM. LOC.	
29	NUMERO OFICINA	NUM. OFC.	
30	OFICINA	OFC.	
31	PANAMERICANA		X

32	PASAJE	PJE.	
33	PEATONAL	PTNAL.	
34	NUMERO DE PISO		X
35	PARQUE INDUSTRIAL	PAR.IND.	
36	PLANTA BAJA	PB	
37	PUEBLO		X
38	RECINTO	RCTO.	
39	SECTOR		X
40	SENDERO	SEND.	
41	SIN NUMERO	S/N	
42	SIN NOMBRE	SN	
43	SUPERMANZANA	SMZ	
44	TRONCAL		X
45	TORRE		X
46	URBANIZACIÓN	URB.	
47	VÍA		X
48	ZONA		X

El procedimiento a realizar, para la estandarización consiste en buscar las diferentes combinaciones de abreviaciones, lo cual no es un proceso automático sino más bien manual, por ejemplo, para AVENIDA se puede tener:

- AV
- AVE.
- AVE
- AVEN.
- AVEN
- AVNDA
- AVNDA.
- AVENDA.
- AVENDA
- AVENIDA

Como se puede ver el proceso es casi predecir lo que el usuario puede considerar como una abreviación válida, por este motivo la estandarización es un tanto tediosa, pero es lo que se realiza ya que el fin es obtener una mejor calidad de la información.

ANEXO 4

Plan de Inconsistencias con cruce de Variables

Directorio de Empresas y Establecimientos

PLAN DE INCONSISTENCIAS CON CRUCE DE VARIABLES

OBJETIVO

Definir las dependencias que existen entre variables que constan en la base de datos a trabajar.

EMPRESAS

ACTIVIDAD ECONÓMICA EMPRESA

La descripción de la actividad debe corresponder a su código en CIIU4

CÓDIGO TELÉFONO

El código telefónico debe corresponder a su respectiva provincia, para los teléfonos que carecen de código de provincia.

EMPLEADOS

Comprobar en empresas que la suma del personal afiliado sea igual a la suma del personal afiliado hombres y mujeres.

Empleados Hombres + Empleados Mujeres = TOTAL

Comprobar en cuanto a unidades locales que la suma de sus empleados sea igual a la que reporta la empresa.

Σ U. Locales = Total Empresa

REMUNERACIONES

Comprobar en empresas que la suma de remuneraciones sea igual a la suma de remuneraciones de hombres y de mujeres.

Remuneraciones Hombres + Remuneraciones Mujeres = TOTAL

VENTAS EMPRESAS

Análisis Comparativo con revistas que publican el ranking de las 100 mejores empresas con nuestra base.

- *Solicitar información al Banco Central de las Exportaciones y las exportaciones.*

ESTADO

Estado pasivo.- Debe tener Fecha inscripción, Fecha de inicio, Fecha de cese.

Estado activo.- Debe tener Fecha inscripción, Fecha de inicio / en el caso de haber cerrado Fecha cese, Fecha de reinicio

TIPO DE UNIDAD LEGAL

Personas Naturales.- No deben tener abreviaturas.

Estructura del RUC

- R.U.C. Jurídicos y extranjeros sin cédula

17	CÓDIGO DE LA PROVINCIA DE EMISIÓN DEL DOCUMENTO
9	**SIEMPRE 9 (FIJO)**
001167	CONSECUTIVO
4	DIGITO VERIFICADOR
001	PRINCIPAL O SUCURSAL

001 corresponde a la sociedad, con su respectiva razón social y otras variables y al local principal o matriz.

- R.U.C. Públicos

17	CÓDIGO DE LA PROVINCIA DE EMISIÓN DEL DOCUMENTO
6	**SIEMPRE 6 (FIJO)**
001167	CONSECUTIVO
4	DIGITO VERIFICADOR
0001	PRINCIPAL O SUCURSAL

El RUC de las entidades públicas tiene siempre el número 6 como tercer dígito, ya pertenezcan al gobierno central, a los gobiernos seccionales municipios y consejos provinciales u otras entidades autónomas provinciales o cantonales a organismos de la función judicial o de la legislativa.

- R.U.C. Persona Natural

17	CÓDIGO DE LA PROVINCIA DE EMISIÓN DEL DOCUMENTO
1	**MENOR A 6 (0, 1, 2, 3, 4, 5)**
001167	CONSECUTIVO
4	DIGITO VERIFICADOR
0001	PRINCIPAL O SUCURSAL

El RUC de una persona natural será de 13 dígitos, sin letras ni caracteres especiales, de los cuales los 10 primeros serán de la cédula de identidad.

Los 2 primeros dígitos corresponden a la provincia donde fue expendida.

Análisis de empleados, sociedades y naturales.

Análisis de la forma jurídica:

- Personas naturales.
- Sociedades.
- Sociedades sin fin de lucro.
- Sector público.

ESTRATOS

Obligados a llevar contabilidad están dentro de los estratos altos.

CONTRIBUYENTES RISE

Sus ventas deben ser menores a cierto valor (definir tope).

UNIDAD LOCAL

Toda empresa por lo menos debe tener una unidad local. Por lo menos un establecimiento este abierto para que la empresa esté abierta. En establecimientos no hay ventas.

Revisar la Actividad Económica predominante por provincia, para analizar la Actividad Económica de las Empresas en esa provincia.

Analizar qué actividades están dentro de hogares, sección T (CIIU4).

OBSERVACIONES

Validar que el número de empleados de unidad local sea igual al número de empleados de empresas.

Conclusiones:

El documento ha recopilado los procesos de validación, limpieza y conteos de las variables de la base de datos del DIEE, validaciones que han nacido a partir de la experiencia y el trabajo que se viene realizando en el directorio de empresas, podemos concluir que las validaciones a seguir, y todos los procesos que se detallan en el presente documento serán de gran ayuda para los futuros trabajos que vamos a tener en el Directorio de Empresas, de esta manera la información que el Directorio publica es una información veraz, y los procesos que conllevan a tener esta información serán cada vez más ágiles y automáticos.

Recomendaciones:

Se recomienda tener siempre en cuenta todas las validaciones que tiene el documento, aquí se puede encontrar como proceder de manera correcta al momento del trabajo en la base de datos antes de subir la información a la base del DIEE.

Si en el futuro se generan nuevas validaciones, se recomienda documentarlas, para poder agregarlas a este documento y así tener siempre un Plan de Validación y Tabulación actualizado.

Fecha de elaboración: 27 de octubre de 2017		
Elaborado por:	Carolina Mina	
Aprobado por:	Libertad Trujillo	