

El Reto del Aseguramiento de la Calidad de los Datos

Una de las tareas claves de las áreas de tecnología, es asegurar la calidad de los datos, en un entorno donde la convivencia de más de un sistema operativo es algo habitual, donde existen múltiples bases de datos y donde la generación de diversos formatos de ficheros deben ser integrados en entornos homogéneos para poder explotarlos adecuadamente.

Las áreas de tecnología deben asegurar la exactitud, totalidad, oportunidad, relevancia, consistencia y nivel de detalle de los datos para atender las distintas necesidades de negocio, distintos niveles de usuarios; un verdadero reto.

Los responsables de atender esta necesidad de auditar en forma permanente los datos para asegurar su calidad; deben resolver no pocos problemas.

El flujo de la información en una organización es altamente dinámico y la toma de decisiones con esa información depende en gran medida de una correcta auditoría de datos.

La Solución Dynamic Data Web

Análisis de Calidad y Auditoría Dinámica de Datos

Dynamic Data Web dispone de una serie de características que permiten implantar un sistema de auditoría dinámica de datos.

Esta auditoría dinámica brinda la posibilidad de cargar e integrar datos desde múltiples fuentes, con velocidades de carga de 10 Gb. por hora de media.

Este veloz proceso de carga, además indexa todos los datos sin ocupar más espacio que en el origen generando un guardado dinámico en un repositorio analítico basado en columnas.

Inmediatamente a continuación de este proceso, se pueden comenzar el proceso de auditoría, sin la necesidad de construir modelados ni aplicar complejos algoritmos.

Gracias a su entorno visual, su guardado en máxima granularidad, exploración, herramientas de análisis y componentes de transformación, Dynamic Data Web permite realizar sobre la marcha los procesos que determinan el estado de la calidad de los datos y su corrección.

- Dynamic Data Web dispone de un entorno de **Exploración**, creación de grupos de datos, aplicación de condiciones lógicas, ordenación y muestreo, para visualizar y operar con los datos con toda profundidad.
- Dynamic Data Web incorpora un conjunto de componentes de **Análisis Dinámico** como el análisis concurrencia por diagramas de Venn, de múltiples cruces por Pivot Tables, de características identificativas mediante Profile, de agrupaciones por Bubble, etc.; que aportan una visión altamente flexible para tomar decisiones sobre los datos.
- Dynamic Data Web, cuenta con un set de componentes de **Herramientas de Ingeniería**, para generar transformaciones y correcciones, a través de decodificaciones, tramificaciones, rangos, selección, métricas basadas en funciones, así como agregados y sumalizaciones.

Estas características en conjunto, reportan grandes beneficios a las áreas de tecnología y les permite optimizar sus recursos y asegurar la máxima calidad de datos de sus sistemas.

Ejemplos de Auditoría Dinámica de Datos en Acción

Compleitud

Con las funcionalidades de exploración dinámica, Dynamic Data Web permite detectar rápida y visualmente la completitud de los campos de una tabla.

La funcionalidad de sumario de una tabla, presenta un grid de datos, donde por cada campo podemos ver su tipología, valores discretos, nulos y la relación entre valores discretos respecto del total junto con la relación de registros nulos respecto del total.

Esta primer visión, permite tener panorama del estado de la calidad de los datos e incidir en los campos de mayor relevancia para iniciar el proceso de auditoría, explorando los registros y campos en todo su nivel de detalle para visualizar su contenido y tomar acciones correctivas.

Estandarización y Consistencia de Datos

El entorno de exploración de Dynamic Data Web, permite operar con cientos de millones de registros y con muy poco hardware, de forma visual e inmediata.

La sola selección de un campo de una tabla con el ratón, nos presenta los valores discretos que toma ese campo, una curva de distribución y los parámetros estadísticos habituales, para el caso de campos numéricos.

Un ejemplo de la verificación de la estandarización y consistencia, se puede presentar con el campo de edad o el de sexo.

El campo de sexo que toma más de 2 valores discretos (por ejemplo, mujer, hombre, empresa) indica en principio falta de consistencia.

A través de funciones de selección, podemos visualizar el detalle de los registros no consistentes y explorar dinámicamente hasta llegar a una conclusión que permita aplicar una regla de transformación.

Si vemos los valores discretos en contenido, puede indicar falta de estandarización. Por ejemplo, si el campo sexo toma 4 valores y estos son femenino, masculino, f y m, es necesario estandarizar. Utilizando la funcionalidad de decodificación se genera un nuevo campo estandarizado.

Un caso similar se presentaría con la exploración del campo de edad, por ejemplo si tomase valores excesivos o negativos.

Detección de Duplicados

A través de la utilización de exploración de campos y la presentación de valores únicos y visualización al detalle, Dynamic Data Web permite detectar duplicidades visualmente.

Un ejemplo es la identificación única de una persona o un expediente en un registro maestro.

Con el solo posicionarnos en una tabla, en el campo que corresponda al identificador único y acceder a la opción de valores únicos y registros, verificamos su coincidencia y la existencia de duplicidad.

Un registro 1.000.000 de ciudadanos debe contener 1.000.000 de valores diferentes, en caso contrario existe una duplicación o una inconsistencia.

Utilizando las funciones de selección visual, se pueden arrastra y soltar el grupo de datos duplicados y visualizarlos al detalle. Podría suceder que existan varios registros del mismo ciudadano (duplicación) o registros de distintos ciudadanos con el mismo identificador (inconsistencia).

Mediante la utilización de las Herramientas de Ingeniería, se realiza la corrección para asegurar la calidad de los datos.

Integridad y Linkado de Datos

A través de la exploración y visualización de datos en forma dinámica, junto con herramientas de análisis; Dynamic Data web permite detectar y corregir problemas de integridad de datos, registros huérfanos y relaciones no esperada.

Un ejemplo es la verificación de la integridad de datos de ciudadanos y la petición de un tipo de subvención. No deberá haber más subvenciones que ciudadanos, puesto que no serían íntegros los datos.

En el mismo ejemplo, todas las subvenciones deben tener un ciudadano asignado, de otra forma habría registros huérfanos.

Los análisis de concurrencia detectan estas situaciones en forma visual, y con la misma filosofía permiten seleccionar el grupo de registros no íntegros o huérfanos, para aplicar una regla de transformación y realizar la corrección correspondiente.

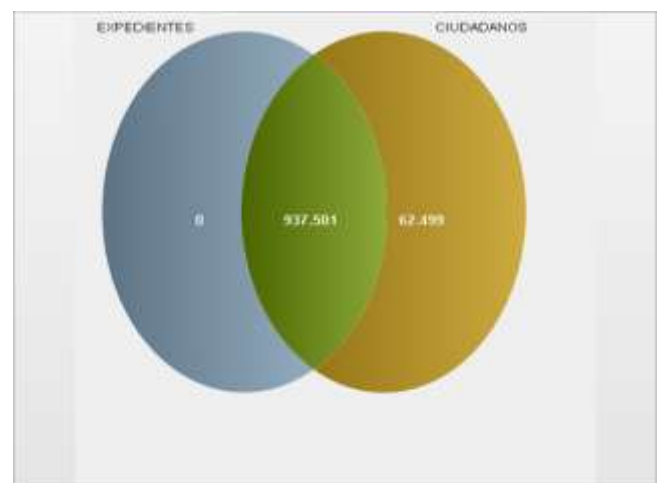
Ejemplos con Datos Simulados de Análisis de Calidad y Auditoría Dinámica de Datos

Análisis de Estandarización y Consistencia



Detección de Falta de Estandarización en el Capo Sexo de una Tabla de Registro de Ciudadanos.

Análisis de Concurrencia y Detección de Huérfanos



Detección de Expedientes Huérfanos, sin asignación relacionada con ningún ciudadano; mediante la comprobación de las tablas de Ciudadanos y Expedientes.

Análisis de Completitud Mediante Presentación de Sumarización

| Nombre | Tipo | Tamaño | Tipo C | Num. Discretos | Nulos | Registros | % Discretos | % Discretos | % Nulos | % Nulos | Index |
|---------------|-------|--------|--------|----------------|---------|-----------|-------------|-------------|---------|---------|-------|
| Id Ciudad | Field | 10 | | 1.000.001 | 0 | 1.000.000 | 100,00 | | 0,00 | | |
| Genitricio | Field | 2 | | 2 | 0 | 1.000.000 | 0,00 | | 0,00 | | |
| Iniciales | Field | 4 | | 463 | 0 | 1.000.000 | 0,04 | | 0,00 | | |
| Apellido | Field | 20 | | 95.836 | 0 | 1.000.000 | 9,58 | | 0,00 | | |
| Nombre ... | Field | 20 | | 759.380 | 0 | 1.000.000 | 75,93 | | 0,00 | | |
| Direccion1 | Field | 64 | | 2.122 | 0 | 1.000.000 | 0,21 | | 0,00 | | |
| Direccion2 | Field | 64 | | 1.984 | 499 | 1.000.000 | 0,19 | | 0,04 | | |
| Direccion3 | Field | 64 | | 6.306 | 0 | 1.000.000 | 0,63 | | 0,00 | | |
| Ciudad | Field | 20 | | 2.352 | 499 | 1.000.000 | 0,23 | | 0,04 | | |
| Provincia | Field | 20 | | 112 | 11.282 | 1.000.000 | 0,01 | | 1,12 | | |
| Codigo P... | Field | 10 | | 532.455 | 0 | 1.000.000 | 53,24 | | 0,00 | | |
| Sexo | Field | 10 | | 3 | 7.212 | 1.000.000 | 0,00 | | 0,72 | | |
| Edad | Field | 2 | | 75 | 0 | 1.000.000 | 0,00 | | 0,00 | | |
| Renta | Field | 8 | | 2.119 | 0 | 1.000.000 | 0,21 | | 0,00 | | |
| Numero H... | Field | 2 | | 7 | 0 | 1.000.000 | 0,00 | | 0,00 | | |
| Fecha Na... | Field | 8 | | 20.922 | 0 | 1.000.000 | 2,09 | | 0,00 | | |
| Permite e... | Field | 2 | | 3 | 0 | 1.000.000 | 0,00 | | 0,00 | | |
| Estado Civil | Field | 10 | | 5 | 0 | 1.000.000 | 0,00 | | 0,00 | | |
| Propiedad... | Field | 10 | | 3 | 222.084 | 1.000.000 | 0,00 | | 22,20 | | |
| Canal Pre... | Field | 10 | | 6 | 0 | 1.000.000 | 0,00 | | 0,00 | | |
| Tipo Vivie... | Field | 10 | | 6 | 36.149 | 1.000.000 | 0,00 | | 3,61 | | |
| Profesion | Field | 16 | | 12 | 0 | 1.000.000 | 0,00 | | 0,00 | | |

Presentación en Resumen de cada uno de los campos de una Tabla

Análisis Comparativo y Detección de Duplicados

| Registros | Sumario | Valores Discretos | Gráfico | Estadísticas | Frecuencia |
|------------|---------|-------------------|---------|--------------|------------|
| 1 / 10001 | | | | | |
| Valor | Número | | | | |
| 1549700391 | 9 | | | | |
| 1549700524 | 3 | | | | |
| 1549700631 | 2 | | | | |
| 1549700383 | 1 | | | | |
| 1549696979 | 1 | | | | |
| 1549700011 | 1 | | | | |
| 1549700342 | 1 | | | | |
| 1549700763 | 1 | | | | |
| 1549701969 | 1 | | | | |
| 1549702256 | 1 | | | | |
| 1549702264 | 1 | | | | |
| 1549701639 | 1 | | | | |
| 1549700870 | 1 | | | | |
| 1549701332 | 1 | | | | |
| 1549701605 | 1 | | | | |
| 1549694776 | 1 | | | | |
| 1549694990 | 1 | | | | |
| 1549695617 | 1 | | | | |
| 1549694701 | 1 | | | | |
| 1549694495 | 1 | | | | |
| 1549694511 | 1 | | | | |
| 1549694602 | 1 | | | | |
| 1549696086 | 1 | | | | |

| Registros | Sumario | Valores Discretos | Gráfico |
|-----------|---------|-------------------|---------|
| 1 / 10001 | | | |
| Valor | Número | | |
| H1857752W | 1 | | |
| H1857780G | 1 | | |
| H1857904L | 1 | | |
| H1857552E | 1 | | |
| H1857467P | 1 | | |
| H1857471C | 1 | | |
| H1857535W | 1 | | |
| H1857933S | 1 | | |
| H1858436D | 1 | | |
| H1858451L | 1 | | |
| H1858477S | 1 | | |
| H1858317J | 1 | | |
| H1858002M | 1 | | |
| H1858048L | 1 | | |
| H1858212D | 1 | | |
| H1855739O | 1 | | |
| H1855770O | 1 | | |
| H1855775E | 1 | | |
| H1855655P | 1 | | |
| H1855242T | 1 | | |
| H1855335R | 1 | | |
| H1855437C | 1 | | |
| H1855820N | 1 | | |

Comparación de Registros de DNI y Número de Identificación de una Tabla de Ciudadanos donde se visualiza la existencia de más de un número de identificador para un mismo DNI.