Implementación de un Data Mart para una gerencia de la Agencia de Recaudación Tributaria de Río Negro.

Rossana Ayelén Traiman Schroh[1],Sonia Formia[2] y Nicolás García Martínez[2]

[1] Universidad Nacional de Río Negro
{ratraimanschroh}@unrn.edu.ar
[2] Laboratorio de Informática Aplicada
Universidad Nacional de Río Negro
{sformia,ngarciam}@unrn.edu.ar

Resumen. Muchas organizaciones del Estado se encuentran en proceso de incorporar paulatinamente la ciencia de datos, en alguna de sus formas, como medio para brindar cada vez mejores herramientas a los funcionarios encargados de la toma de decisiones y el seguimiento de las políticas de estado. En este sentido la Agencia de Recaudación Tributaria de la Provincia de Río Negro ha desarrollado un Data Warehouse para el manejo de la información tributaria que ha evolucionado desde el año 2012 a la actualidad. En este trabajo se presenta la creación de un Data Mart, integrado a dicho Data Warehouse, para abordar la información sobre los Recursos Humanos de la organización, siguiendo el enfoque Temático definido por Bill Inmon.

Palabras claves: Inteligencia de negocios, Data Warehouse, Data Mart, Soporte a la Toma de Decisiones.

1 Introducción

1.1 Contexto

El presente trabajo describe el avance realizado durante implementación de un Data Mart integrado al Data Warehouse (DW) de la Agencia de Recaudación Tributaria de Río Negro (ARTRN) y, específicamente destinado a la Gerencia de Recursos Humanos (RRHH) de la misma.

La ARTRN cuenta con la implementación de un DW desde el año 2012 que está compuesto por diferentes Data Marts (DM), cada uno de ellos destinado a diferentes áreas dentro del organismo, pero principalmente orientados al ámbito tributario y son:

RECAUDACIÓN - Contiene información histórica de la recaudación de impuestos.

CUENTA CORRIENTE - Información detallada e histórica de las obligaciones fiscales de cada contribuyente los saldos y pagos.

DECLARACIONES JURADAS - Información detallada de las declaraciones realizadas por los contribuyentes de los diferentes tributos.

RETENCIONES Y PERCEPCIONES - Incluye información en detalle de las retenciones y percepciones tanto de lo declarado por el contribuyente como la que informan los Agentes de Recaudación/Percepción.

DEUDA GESTIONADA - Información de Planes de Pago, Intimaciones, Juicios que influyen en la gestión de la deuda.

MONITOREO USUARIOS / TRÁMITES / MESA DE AYUDA - Contiene información sobre la gestión de usuarios y trámites tanto externos como internos junto a la efectividad en su tratamiento [6].

Debido a la gran cantidad de información que se almacena en las bases de datos operacionales de la Agencia, fue necesaria la creación de esta herramienta para poder realizar análisis de los mismos de forma integral, histórica y de fácil manipulación.

Sin embargo, como se mencionó anteriormente, la mayor parte del DW está orientado al objetivo principal del Organismo que es la recaudación tributaria. Por lo que desde la Gerencia de RRHH no contaban con esta herramienta para poder analizar la información que le compete al área.

2 Delimitación del problema

2.1 Estado inicial

La Gerencia de Recursos Humanos cuenta principalmente con un sistema web desde donde se administra y gestiona todo lo referente a los datos del personal que

pertenece a la ARTRN. Este sistema almacena la información en una base de datos operacional.

Por otro lado cuentan con un archivo en formato Access, administrado mediante la herramienta Microsoft Access en donde almacenan información adicional y les sirve como base para realizar informes.

Si bien ambas herramientas permiten realizar una serie de reportes, están muy limitados en cuanto al tipo de información que pueden extraer de los mismos. Si se necesita por ejemplo cualquier cambio sobre los informes que están realizados sobre el sistema web inevitablemente tiene que intervenir la Gerencia de Tecnologías de la Información(GTI), debiendo solicitar el cambio, esperar que sea desarrollado, probado e implementado. También, cuando necesitan un reporte totalmente nuevo, deben recurrir a la misma GTI teniendo que esperar a que se cumpla todo el ciclo de desarrollo de la consulta.

Otra limitación es el hecho de que al tener dos lugares diferentes donde almacenar información, cada uno de ellos brinda informes independientes que deben ser descargados, cada uno como una hoja de cálculo, y el cruce de esa información se hace de forma manual. Esto puede generar errores e inconsistencia de datos difícil de detectar.

3 Implementación del Data Mart

Para lograr solucionar estos inconvenientes, se decidió aplicar una herramienta de Inteligencia de Negocios, creando un Data Mart para dar soporte al análisis en la toma de decisiones de la Gerencia de Recursos Humanos.

Los recursos tecnológicos que se utilizaron fueron los que ya estaban disponibles en la ARTRN y siendo usados por el resto de los DM que conforman el DW de la Agencia y son:

- Toad for Oracle Xpert: Utilizado para la administración y gestión de la base de datos.
- Suite Pentaho Enterprise: Es una plataforma compuesta por diferentes herramientas que ofrece soluciones propias, tanto para desarrollar como para mantener y exportar un proyecto de BI. La versión utilizada es la 8.2 Release 8.2.0.0-324 y los componentes del conjunto de herramientas son los siguientes:
    - Pentaho Data Integration (PDI) es la herramienta que permite realizar tareas y procesos ETL
    - Pentaho Interactive Reporting: herramienta de reporting interactivo.
    - Pentaho Analyzer: Es la solución para explotar cubos multidimensionales, gestionado por un motor Mondrian, que es un servidor para el procesamiento analítico.
    - Pentaho Dashboards: permite la construcción de tableros de control o de mando en la interfaz web.

- Pentaho Enterprise Console: herramienta web que permite la configuración, administración y monitoreo de la plataforma.
- Schema Workbench: Permite la definición de los cubos y es donde se plasma la auto documentación para los usuarios finales.

3.1 **Metodología de trabajo**

La implementación del DM se realizó siguiendo la metodología conocida como Bottom-up [1]. Se llevó a cabo una sola iteración desarrollando cada una de las cinco etapas propuestas por esta metodología y descritas a continuación.

3.1.1 Definición del modelo lógico del negocio

Para poder obtener la definición del modelo lógico de negocios se realizaron diversas entrevistas con personal de la Gerencia de RRHH. El resultado de cada una quedó planteado de forma que se pueden identificar: el objetivo de la reunión, y el resultado o resumen obtenido de la misma.

A partir de las reuniones se logró identificar los objetivos particulares de análisis que serían funcionales al área.

3.1.2 Definición del modelo lógico del Data Mart

Con los objetivos ya identificados, se determinan las medidas, dimensiones y la granularidad de cada una de ellas. También se acordó la realización de un único cubo en esta instancia que contenga lo definido anteriormente.

Se definieron las siguientes medidas [5]:

- Cantidad: Hace referencia a la cantidad de empleados. La mayoría de los reportes que se necesitan para el análisis responden a las preguntas del tipo: “¿Cuántos empleados cumplen determinadas condiciones?”

- Edad: Esta medida a diferencia de la anterior, es de tipo no aditiva debido a que no tiene sentido sumarla. Responde por ejemplo a la siguiente solicitud: “empleados activos mayores a 60 años”.

Por otro lado, como dimensiones se plantearon [5]:

- Fecha de foto: En primer lugar se considera la dimensión tiempo para garantizar la perspectiva de almacenamiento histórico que tiene la información, y a la que se debe poder acceder para el análisis de tipo evolutivo de la planta de empleados de la Organización. Esta dimensión está conformada por una jerarquía con tres niveles: Año, Mes y Día.

- Empleados: Hace referencia a los datos particulares y personales de cada empleado de la Agencia como: apellido y nombre, número de CUIL/CUIT, entre otros.

- Categoría: Esta dimensión se conforma como una jerarquía con niveles que hacen referencia a las clasificaciones mediante escalafones dentro de la Organización. Los tres niveles identificados nombrándolos de menor a mayor detalle son: Régimen, Agrupamiento y Categoría.

- Estado civil: Permite clasificar a los agentes por estado civil.

- Nacionalidades: Esta dimensión se crea para brindar información sobre las nacionalidades de los empleados.

- Unidad organizativa: Esta dimensión se refiere a la sede de funciones del empleado que puede ser por ejemplo una subdelegación, receptoría, gerencia, entre otras.

- Organización: Esta dimensión indica el Área/Gerencia donde se ubica el empleado, por ejemplo Tecnologías de la Información.

- Régimen jubilatorio: Para algunos reportes es necesario conocer si un empleado está o no jubilado, por lo que se creó esta dimensión.

- Situación de Revista: Se definió la dimensión que identifica a los agentes como por ejemplo en: Planta Permanente.

- Grado Académico: Indica el mayor grado académico obtenido por el agente. Se define el nivel Título, como siguiente nivel se encuentra Tipo de Título y, finalmente se define la Institución en donde el empleado alcanzó su máximo grado académico

- Carga horaria: Especifica si el agente cuenta con carga horaria adicional.

- Licencia sin goce de haberes: Esta dimensión indica si el empleado cuenta con una licencia sin goce de haberes.

- Baja Provisoria o definitiva: Con esta dimensión se puede saber si el agente cuenta con una baja en proceso o provisoria o si tiene una baja definitiva.

- Género: La dimensión indica el género del empleado.

- Localidad: Esta dimensión es de dos niveles e indica la localidad en donde cumple sus funciones el agente y en otro nivel la provincia a la que corresponde dicha localidad.

- Estado: Esta dimensión sirve para clasificar a los empleados en Activos o Dados de Baja.

3.1.3 Definición del modelo dimensional

En esta etapa se analizó la mejor forma de estructurar los datos en cuanto a la tabla de hechos y su relación con las dimensiones. El modelo que se eligió fue el

denominado Modelo Estrella que consiste en tener una única tabla de hechos compuesta por las medidas y una clave foránea a cada una de las dimensiones definidas en la etapa anterior.

3.1.4 Definición del modelo físico

Todo lo que se definió en las etapas anteriores se traduce a un diseño físico en esta fase para poder ser plasmado en una base de datos.

Como la ARTRN ya contaba con un DW productivo, todo lo referente al nuevo DM se integró a esa estructura de base de datos y servidores ya existentes.

Se definió que cada tabla de dimensión tenga un campo de clave primaria y además índices que aceleren la búsqueda de los datos en cada una de ellas.

3.1.5 Proceso de ETL

Todo este proceso de Extracción, Transformación y Carga (ETL, por sus siglas en inglés) se realizó con la herramienta Pentaho Data Integration (PDI) con la que también contaban en la ARTRN. Ésta es una aplicación de escritorio y permite por medio de estructuras que llama Transformaciones (transformations) o Trabajos (jobs) estructurar el flujo de trabajo. Este flujo se determina visualmente en la herramienta por una serie de pasos o steps que están unidos entre ellos a través de flechas de dirección y pueden seguir un camino de éxito o uno o varios alternativos en caso que alguno falle.

Los datos que se extraen desde las fuentes orígenes se hacen de dos maneras. La primera es desde la base de datos operacional del sistema de legajo electrónico y se realiza por medio de vistas que se encuentran en la base de datos del área intermedia o de staging. La segunda fuente origen es el archivo access, que para poder utilizarlo se lo sube al servidor también con steps de la herramienta PDI, y se extrae la información con otros steps definidos para este propósito.

Todo el proceso ETL está estructurado por un job principal, como se puede observar en la Figura 1, que organiza el flujo de tareas dentro del mismo y se puede dividir en tres grandes procesos:

- El primero realiza la búsqueda del archivo Access mediante una conexión FTP y lo copia al servidor para poder trabajarlo desde ahí sin que durante este proceso, el archivo sea modificado por personal de la Agencia.
- La segunda etapa es la que realiza la actualización de las dimensiones esto implica que se buscan desde las fuentes orígenes los posibles datos nuevos cargados desde la última actualización y se dan de alta en cada tabla de dimensión definida.
- La última parte es la que obtiene por un lado, los datos de los empleados del sistema de legajo electrónico a través de una vista creada en el área de Staging. Y por el otro, la información cargada en el Access.

Se lleva a cabo también la tarea de transformación de los datos realizando una limpieza y homogeneización de los mismos. Otra labor que se realiza es el cruce de datos con las dimensiones, reemplazando en cada caso el valor de la dimensión que viene en la staging o Access, por la clave primaria de la misma. De esta forma el último paso que se realiza es el de la carga de datos de la tabla de hechos, en donde se da de alta cada empleado con las medidas seleccionadas y la clave foránea a cada una de las dimensiones.

3.1.5.1 Definición del cubo para la visualización por medio de la herramienta web de Pentaho

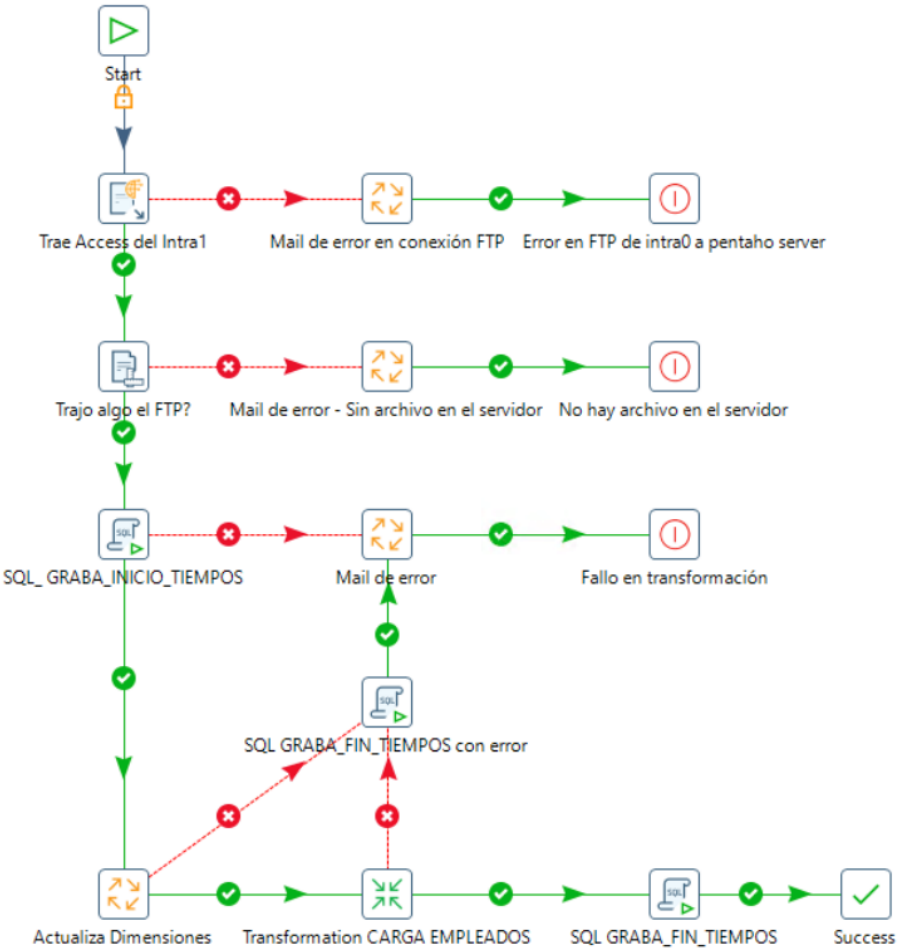

Fig. 1. Job principal del proceso de ETL

Para que los usuarios finales puedan visualizar la información cargada en el cubo, es necesario que se defina la estructura que será interpretada por la herramienta de usuario final, que en este caso es la herramienta web Pentaho Analyzer.

Esta definición se denomina metadata y se realiza con otro componente de Pentaho llamado Schema Workbench, que trabaja sobre un archivo con formato xml. Sobre este archivo se definen las jerarquías de las dimensiones, se formatean las medidas y se detallan los nombres y documentación que será interpretada por el motor de análisis de Pentaho y mostrada a los usuarios.

Mediante esta herramienta además se define como último paso los permisos que tendrán los usuarios para visualizar la información del cubo.

Una vez creada y publicada la metadata en el servidor, los usuarios ya disponen de los datos para comenzar a realizar reportes, permitiéndoles crear reportes de análisis y tableros de control.

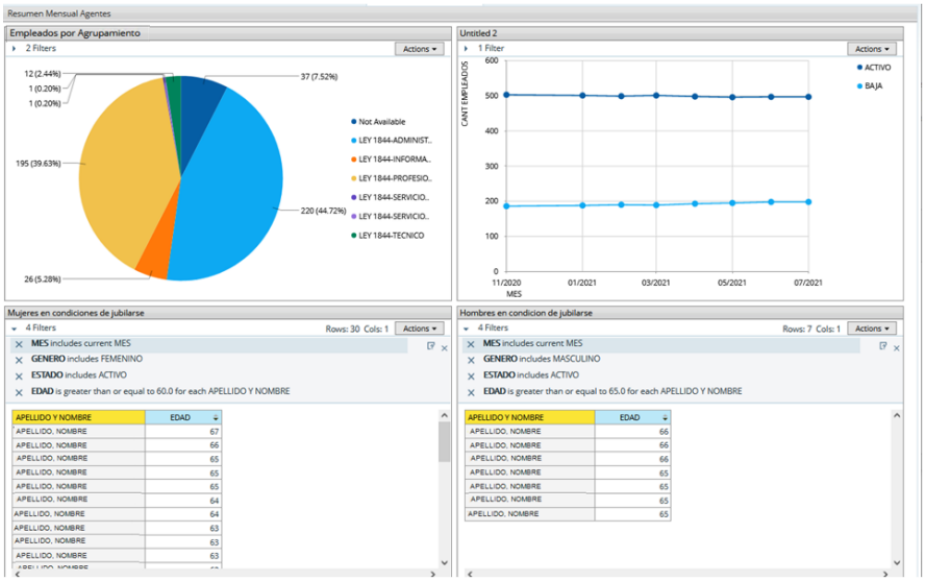

Fig. 2. Tablero de control resultante de reportes sobre el cubo.

4 Conclusiones y trabajos futuros

Debido a que todo el proceso implementado realiza de forma automática el cruce de información entre las dos herramientas de almacenamiento de información, se evitan los errores que puede haber por hacerlo de forma manual.

El acceso a toda la información también está centralizada y estructurada de forma que pueda ser analizada en su conjunto, teniendo una visión completa de la

información y utilizando el mismo formato, reglas y nomenclaturas en los datos independientemente de los orígenes.

La principal ventaja dentro de la Gerencia de RRHH es que el tiempo de armado y definición de reportes es mucho menor, permitiendo dedicar mayor cantidad de recursos al análisis. Esto es porque se evita que los agentes de dicha gerencia tengan que integrar los datos, formatearlos, controlar la información para evitar errores, etc, de forma manual. Sino que todo esto ya se realiza de forma automática en la carga del cubo.

Con el uso de la herramienta se logran analizar medidas de tendencia, informes históricos y un análisis general para poder identificar si son necesarios cambios que permitan mejorar el trabajo dentro de la Agencia.

En la actualidad los usuarios se encuentran trabajando ya con el Data Mart, lo que genera nuevas necesidades dada la naturaleza típicamente iterativa y modular de este tipo de proyectos, se espera en el futuro el agregado de nuevas dimensiones y medidas al cubo existente y la creación de nuevos cubos con información referida a las licencias anuales, el presentismo y los salarios.

Referencias

1. Ross, M. y Kimball, R..: The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Second Edition). (2002)
2. Bernabeu, R. D: DATA WAREHOUSING: Investigación y Sistematización de Conceptos– HEFESTO: Metodología propia para la Construcción de un Data Warehouse. (2009)
3. Formia, S., y Estévez, E.: Implementación y maduración de un data warehouse – caso de estudio de la Agencia de Recaudación Tributaria de Río Negro (ARTRN). (2017).
4. Inmon, W. H.: Building the Data Warehouse. John Wiley & Sons, Ltd. (2002)
5. Traiman Schroh R. A: Implementación de un Data Mart para la ayuda en la toma de decisiones de la Gerencia de Recursos Humanos. (2021)
6. Formia, S., Estevez, E.: Implementación y Maduración de un Data Warehouse – Caso de Estudio de la Agencia de Recaudación Tributaria de Río Negro (ARTRN) 46JAIIO - SIE - ISSN: 2451-7534