Universidad de la República - Facultad de Ingeniería - Instituto de Computación

Proyecto de Grado

Data Warehouse de Programas Sociales del MIDES

Nicolás Álvarez de Ron
Marcelo Bernasconi
Wilson Goicoechea

Tutores
Adriana Marotta
Flavia Serra

Montevideo, Uruguay
Diciembre 2012

Resumen

Este trabajo surge como respuesta a la necesidad del Ministerio de Desarrollo Social (MIDES) de mejorar su eficiencia en lo que respecta al análisis de la información obtenida por las intervenciones de sus programas sociales. Hoy en día dicha información no está centralizada. Las diversas fuentes de datos existentes están dispersas, y su integración, cuando se puede, se realiza en forma artesanal. Este hecho origina un gran problema en lo que respecta a la calidad de los datos y al tiempo consumido cada vez que se desea analizar dicha información, además de hacer dificultosa la consolidación de la misma.

El objetivo del presente proyecto es la construcción de un prototipo de Data Warehouse corporativo destinado a la explotación de los datos de los programas sociales implementados por la institución. El proceso comienza con la elección de los programas sociales *Uruguay Integra* y *Tarjeta Uruguay Social,* posteriormente, se realizan las reuniones de relevamiento con los usuarios, el análisis de la realidad de ambos programas, el análisis de los datos, el estudio de las herramientas a utilizar y el diseño de la solución. Luego se definen e implementan los procesos de limpieza de datos, se construyen los procesos de carga, se genera la documentación correspondiente a dichas actividades y finalmente se implanta el prototipo en un ambiente de testing accesible a los usuarios para la evaluación del mismo.
El diseño de la arquitectura debió permitir flexibilidad tanto en las herramientas utilizadas para los procesos de carga y limpieza de datos como en las utilizadas para el análisis de la información. El prototipo se implementó utilizando *MS SQL Server*, *Integration Services*, *Analysis Services* y *Excel*. Con el objetivo de evaluar alternativas a dichas herramientas, parte del prototipo fue replicado utilizando *Pentaho Data Integration* e *Ideasoft O3*.

Contenido

1 Introducción

Una correcta gestión de datos es esencial para el buen funcionamiento de las organizaciones. Hoy en día las mismas cuentan, en general, con sistemas formales de información basados en herramientas informáticas. Aun así, cuando la cantidad de datos acumulada es muy grande, resulta difícil identificar aquellos que verdaderamente son relevantes. Se vuelve indispensable contar con sistemas que permitan transformar esos datos en información útil que ayude a tomar decisiones a nivel gerencial, permitiendo que las funciones de planeación y control se realicen de la manera más eficaz. La información debe ser accesible, confiable y fácil de interpretar.
Los sistemas de Data Warehouses surgieron a principios de la década del noventa como una respuesta a esta necesidad, brindando solución a problemas críticos que no podían ser resueltos de forma eficiente por los sistemas de información operacionales, como la integración y calidad de los datos, y el acceso a los mismos.
Los sistemas de información operacionales dan soporte a procesos transaccionales. Sus estructuras de almacenamiento de datos no son las adecuadas para cumplir eficientemente los objetivos del análisis, siendo sus puntos más débiles la ausencia de información histórica y los largos tiempos de respuesta para los tipos de consulta requeridos.

En este proyecto se pretende aplicar los beneficios de los sistemas de Data Warehouse al área social del estado, construyendo un sistema que permita una correcta explotación de los datos de los programas del Ministerio de Desarrollo Social.

A continuación se realiza una presentación del contexto del proyecto. Luego se describe la motivación del problema, los objetivos planteados y los resultados esperados.

1.1 Ministerio de Desarrollo Social

El Ministerio de Desarrollo Social (MIDES) fue creado en marzo de 2005 con el objetivo de instrumentar el Plan de Atención Nacional a la Emergencia Social (PANES). La implementación de dicho plan buscaba garantizar la cobertura de las necesidades básicas de las personas más vulnerables y construir, de manera colectiva, mecanismos de salida de la pobreza e indigencia, en el marco de un proceso de integración social.
La urgente necesidad de poner en marcha los proyectos de reestructura de las políticas sociales, y la escasez de recursos disponibles, relegaron a un segundo plano a los aspectos organizativos, formales e institucionales del ministerio.
Actualmente el MIDES tiene la responsabilidad de diseñar e implementar las políticas sociales nacionales, así como realizar la coordinación, el seguimiento, la supervisión y

evaluación de los programas destinados a favorecer el acceso de los sectores sociales vulnerables a las prestaciones, bienes y servicios sociales.
Los aspectos institucionales previamente relegados hoy pretenden ser fortalecidos. Sin embargo, como consecuencia negativa de la forma de trabajo en los comienzos, se tiene hoy una fuerte descentralización de la información de los programas que el MIDES lleva adelante. La consolidación e integración de datos de las fuentes diversas y heterogéneas existentes se realiza en forma artesanal, lo que hace dificultosa la acción coordinada entre los diferentes programas que además genera una enorme cantidad de inconsistencias que enlentece y hace arduos los procesos de seguimiento y evaluación.
En este último tiempo, la División Informática del ministerio ha tomado un rol protagónico en relación a este tema y ha iniciado la construcción de un sistema que a mediano plazo centralizará los datos correspondientes a las intervenciones de todos los programas sociales.

1.2 Motivación

Durante el año 2009, el equipo de Informática del MIDES trabajó en la elaboración de un nuevo modelo de gestión de los sistemas de información. Se definieron lineamientos básicos referentes a los sistemas informáticos y bases de datos.
Con el apoyo de una consultoría externa, se relevó la situación del área Informática, se realizó un análisis crítico y se elaboró el *Plan Director de Sistemas Informáticos* [1], en el cual se presentaron propuestas para la mejora del área a través de la definición de estándares tecnológicos, de seguridad, de comunicación entre los sistemas y de arquitectura de las aplicaciones. También se propuso una reestructura organizativa del área y se confeccionó un portafolios de proyectos para los años siguientes.
Dentro de esos proyectos se encontraba la construcción de un Data Warehouse destinado a la explotación de los datos de los programas sociales del ministerio. La realización de dicho proyecto representaba una gran oportunidad de brindar herramientas poderosas de análisis para la toma de decisiones.
A mediados del año 2010 se implementó, en varios organismos del estado, el *Tablero de Control Ministerial* (TCM). Este hecho representó un primer acercamiento del ministerio a las herramientas de análisis antes mencionadas.
El TCM es un producto del Área de Gestión y Evaluación del Estado (AGEV). Se trata de un sistema que permite monitorear y gestionar las políticas públicas sobre la base de herramientas de Inteligencia de Negocios. La empresa *Ideasoft* [2] fue la encargada de la implementación de un prototipo de TCM en el MIDES.
Sin embargo, las dificultades presentadas por las fuentes de datos sumadas a la ausencia de procedimientos claros para la integración y limpieza de datos y, en general, para el

mantenimiento del sistema, provocaron el desinterés por su utilización, tanto de los usuarios como de los informáticos. Esto produjo que el proyecto quedara inactivo.

1.3 Objetivos

Como ya ha sido mencionado, la ausencia de un sistema informático centralizado que registre las intervenciones de los programas sociales ha concluido en la existencia de múltiples y diversas fuentes de datos de dificultosa integración para su análisis.
El objetivo de este proyecto es construir un prototipo de sistema de Data Warehouse corporativo que permita el análisis de la información de las intervenciones de los programas sociales del MIDES.
Dentro del alcance del proyecto está prevista la implementación para dos programas sociales: *Uruguay Integra* y *Tarjeta Uruguay Social*. La elección de estos dos programas estriba en su naturaleza diversa, lo que permite, mediante el estudio de sus distintas realidades, abarcar gran parte de la problemática del MIDES.
No obstante, debe establecerse una solución escalable que permita la futura incorporación de los múltiples programas sociales ejecutados por el ministerio.

1.4 Resultados esperados

Con la realización del proyecto de grado se busca obtener:

- Una solución que permita construir un Data Warehouse corporativo con la mayor independencia posible de las herramientas utilizadas para su construcción. Además, dicha solución debe considerar la incorporación de nuevas fuentes de datos y nuevos programas sociales.

- La implementación de un prototipo para el programa social Uruguay Integra.

- La implementación de un prototipo para el programa Tarjeta Uruguay Social.

- La definición e implementación de los procesos de carga y actualización automáticos del Data Warehouse.

- La documentación de la solución.

- La implantación de los prototipos en ambiente de testing.

1.5 Contenido del documento

Este documento contiene 8 capítulos. En el capítulo 1 se presenta una introducción general del proyecto. En el capítulo 2 se explican brevemente los conceptos necesarios para la comprensión del documento. La arquitectura general de la solución y las decisiones de diseño se presentan en el capítulo 3, mientras que en el capítulo 4 se describe el análisis y diseño de la solución para los dos programas sociales implementados. El capítulo 5 trata todo lo referente a los procesos de carga de datos al Data Warehouse construido. Por otro lado, los aspectos relacionados con la implementación del proyecto, los datos y las herramientas utilizadas se describen en el capítulo 6. En dicho capítulo también se muestran algunos ejemplos del resultado mediante imágenes de las diferentes herramientas de análisis. En el capítulo 7 se realiza una evaluación de los resultados alcanzados, las dificultades encontradas y las posibles extensiones al trabajo, como así también se presentan las conclusiones finales. Finalmente, en el capítulo 8 se listan todas las referencias bibliográficas utilizadas.

2 Conceptos básicos

En este capítulo se introduce brevemente el marco teórico del proyecto. Se explican los conceptos considerados relevantes para la comprensión del trabajo realizado.

2.1 Sistemas de Data Warehousing

En la actualidad, la gran mayoría de las empresas posee sistemas informáticos de gestión, los cuales generan una gran cantidad de datos acerca de su actividad. Es deseable que esos datos puedan ser procesados de alguna manera, de forma tal, que puedan ser transformados en información útil destinada al aumento de la productividad de la organización. Sin embargo, aunque los datos se encuentren disponibles, suelen existir problemas que impiden su correcto aprovechamiento, como por ejemplo, el acceso a ellos en forma eficiente, asegurando además su óptima calidad.
Los Sistemas Operacionales de gestión con los que suelen contar las empresas no resultan eficientes en este sentido. El gran volumen de datos demandados para el procesamiento de las consultas de índole gerencial hace que los tiempos de respuesta de estos sistemas aumenten considerablemente, degradando su rendimiento. A este hecho, se le suma la dificultad para almacenar datos históricos e integrar datos heterogéneos de fuentes diversas. Todo esto hace que la mayoría de los datos, con los que cuentan las empresas, comúnmente no sean explotados.
Los Sistemas de Data Warehousing buscan brindar una solución a estos problemas. Esta clase de sistemas forma parte del conjunto de los llamados Sistemas de Soporte a Decisiones (DSS), los cuales, a diferencia de los Sistemas Operacionales, son desarrollados con el objetivo específico de apoyar de forma eficiente el proceso de toma de decisiones gerenciales. Su arquitectura típica se basa en un proceso de integración aplicado a un conjunto de datos fuente, donde la pieza fundamental es el Data Warehouse.

2.2 Data Warehouse

Según la definición clásica de Bill Inmon, un Data Warehouse (DW) es un conjunto de datos orientado a temas, integrado, no volátil y variable en el tiempo, el cual se organiza para brindar soporte a la toma de decisiones [3].
Que el conjunto sea orientado a temas significa que los datos están organizados de forma tal que todos los elementos de datos relativos al mismo evento u objeto de negocio queden relacionados entre sí.

Que sea integrado expresa que el DW es el resultado de la integración de fuentes diversas, como sistemas operacionales o archivos con datos operativos de la organización.
No volátil significa que en un DW los datos son cargados y accedidos para consulta, pero, a diferencia de lo que sucede en los Sistemas Operacionales, nunca son actualizados (en sentido general), ni eliminados.
Variable en el tiempo significa que cada unidad de datos del conjunto está asociada a un período de tiempo específico, lo que permite realizar comparaciones temporales.

2.3 Data Marts

Un DW incluye datos acerca de toda una organización y es utilizado por los usuarios en los niveles altos de gestión con el fin de apoyar las decisiones estratégicas. Sin embargo, estas decisiones pueden ser tomadas en niveles más bajos de la organización, y estar relacionadas con las áreas de negocio específicas, en cuyo caso sólo se requiere un subconjunto de los datos contenidos en un DW. Este subconjunto suele encontrarse en un Data Mart (DM), que tiene una estructura similar a un DW, pero de menor tamaño [4].

2.4 Modelos Multidimensionales

Los Modelos Multidimensionales representan la información como matrices multidimensionales, llamadas *Cubos*, cuyos ejes se denominan *Dimensiones* y sus elementos *Medidas*.
Las *Dimensiones* son variables independientes que corresponden a los criterios de análisis de los datos. Las *Medidas* son variables dependientes que corresponden a los valores analizados y se encuentran en la intersección de las dimensiones.
Por ejemplo: si se quiere analizar las ventas de bicicletas en función de su marca, color y fecha de venta, las dimensiones corresponderán a los criterios de análisis mencionados; es decir, la marca, color y fecha de venta, mientras que la medida será la cantidad de bicicletas vendidas.
Las dimensiones pueden estar organizadas en jerarquías de agregación para representar diferentes niveles de análisis. A cada nivel de una jerarquía se le denomina *Nivel de Agregación*. Siguiendo con el ejemplo anterior, la dimensión correspondiente a la fecha de venta se podría organizar como una jerarquía con los niveles día, mes y año. También pueden existir varias jerarquías para una misma dimensión, a las cuales se les llama *Jerarquías Alternativas*.

Existen operaciones básicas que permiten navegar por la información contenida en un modelo multidimensional. Las mismas son descriptas a continuación, acompañadas de figuras que sirven de ejemplo.

Se considera el cubo de la Figura 2-1, el cual muestra información de la venta de productos en algunas ciudades europeas.

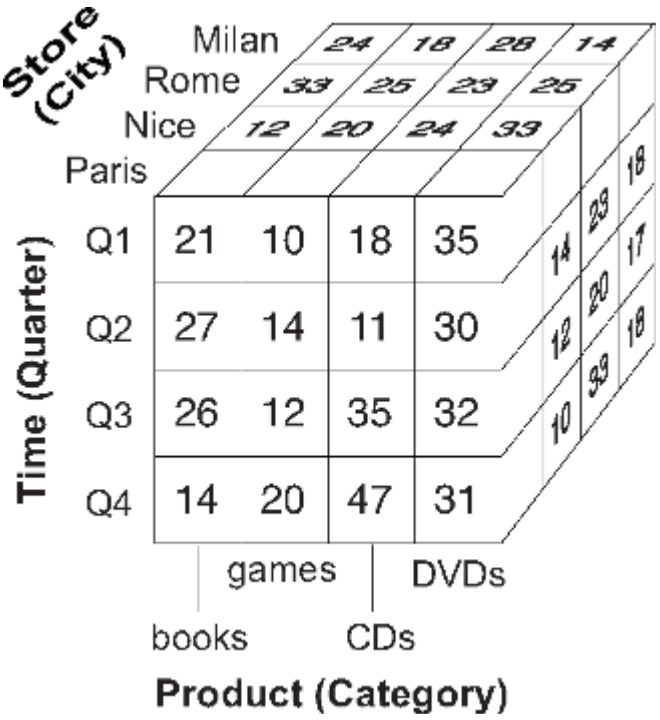


Figura 2-1: Cubo original [4].

Slice

Permite seleccionar un subconjunto de dimensiones sobre los cuales analizar las medidas. Se puede ver un ejemplo en la Figura 2-2, donde se muestra únicamente la información de las ventas de Paris

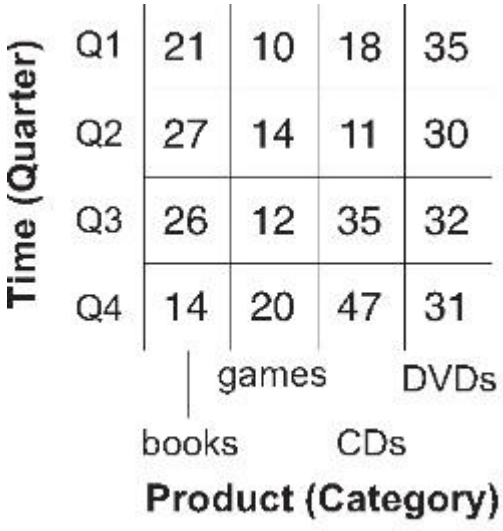


Figura 2-2: Slice en Store City = Paris [4].

Dice

Permite establecer valores fijos para algunas dimensiones. Se presenta un ejemplo en la Figura 2-3, en ella se observa las ventas de París en los dos primeros trimestres.

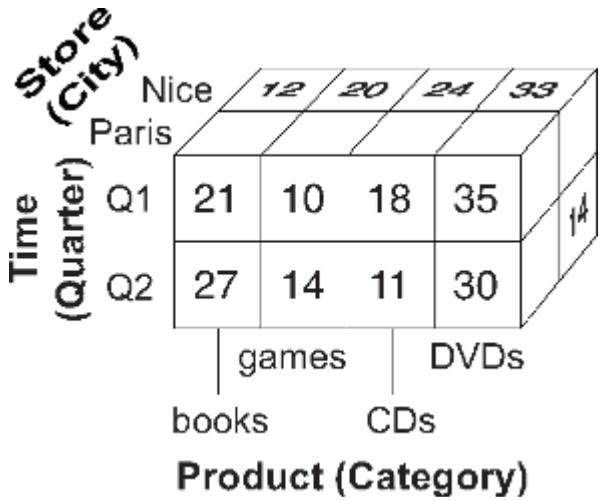


Figura 2-3: Dice en Store.Country=’France’ y Time.Quarter=’Q1’ or ’Q2’ [4].

Rotación

Permite seleccionar el orden de visualización de las dimensiones. Un ejemplo de esta operación se muestra en la Figura 2-4, en la cual se puede observar la información de las ventas por *Tiempo* y *Ciudad*.

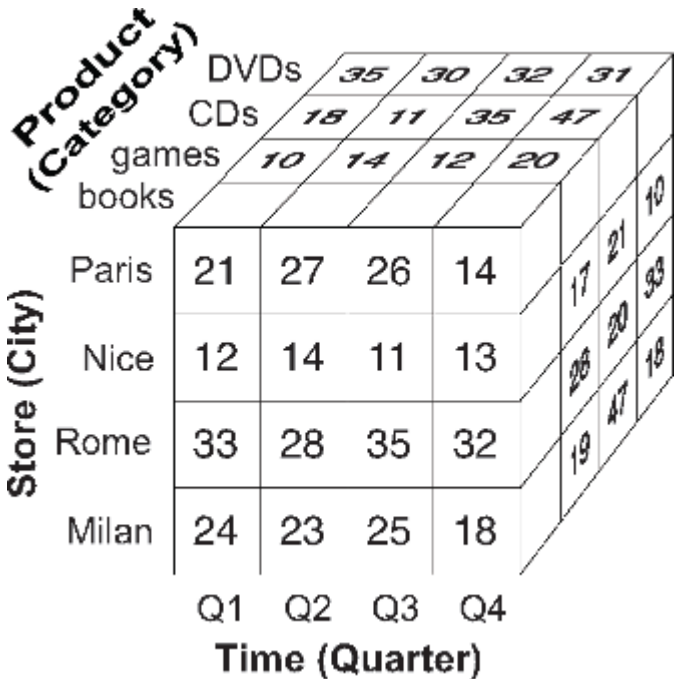


Figura 2-4: Rotación [4].

Drill Up y Drill Down

Permite moverse a través de la jerarquía de una dimensión, agrupándola o desagrupándola respectivamente. En la Figura 2-5 se muestra la operación Drill Down desde el nivel *trimestre* al nivel *mes* en la dimensión *Tiempo.*

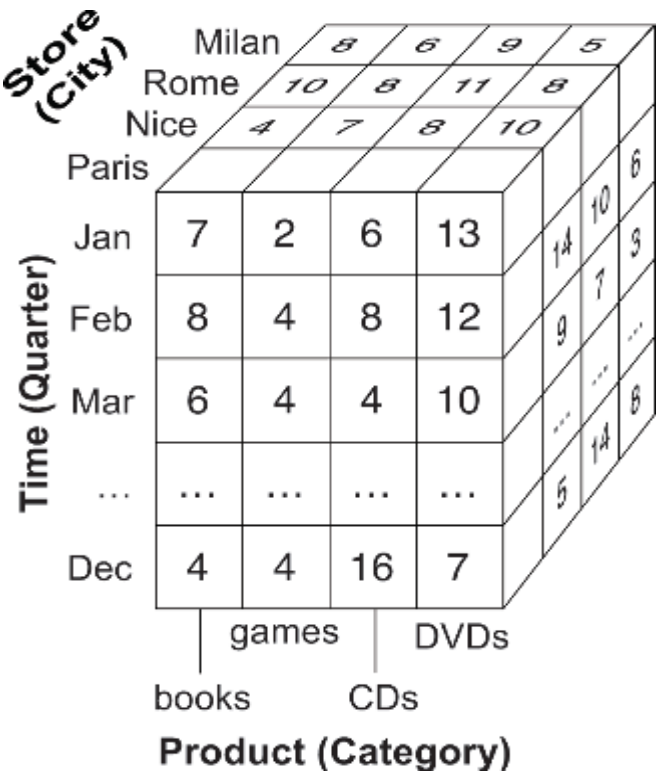


Figura 2-5: Drill down al nivel *mes* [4].

Roll Up

Calcula las medidas en función de los agrupamientos. Se debe especificar cuál es la operación que calcula el valor de la medida agrupada. Dicha operación puede ser, por ejemplo, la suma o el promedio. En la Figura 2-6 se muestra el resultado de aplicar Roll Up, al cubo de la Figura 2-1, desde el nivel *Ciudad* al nivel *País* en la dimensión *Store*. El resultado surge de sumar los valores de las ciudades localizadas en el mismo país.

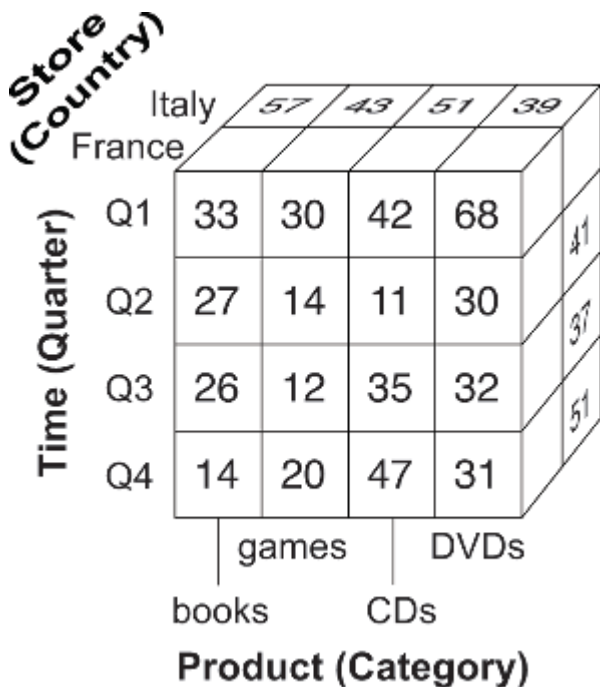


Figura 2-6: Roll up al nivel *país* [4].

2.5 Herramientas OLAP

Los sistemas de Procesamiento Analítico en Línea (OLAP) representan una solución al problema de brindar una respuesta rápida y eficiente a las consultas de grandes cantidades de datos. Las herramientas OLAP extraen los datos almacenados en los DW, o en los DM, y representan la información utilizando estructuras multidimensionales (Cubos). Estas herramientas permiten realizar de forma intuitiva las operaciones vistas en la sección anterior, y están orientadas al usuario de manera tal de brindar una visualización amigable de los datos, por ejemplo mediante cuadros de múltiple entrada o gráficas.

Existen diversas implementaciones de la tecnología OLAP, lo que permite clasificarla en ROLAP, MOLAP u HOLAP, en función del tipo de almacenamiento utilizado.
La implementación ROLAP almacena los datos en bases de datos relacionales, las que son diseñadas de forma adecuada para optimizar la velocidad de las consultas. En cambio, la implementación MOLAP utiliza estructuras multidimensionales para el almacenamiento de los datos. Obtiene un óptimo rendimiento gracias a sus estructuras de almacenamiento específicas, indexación multidimensional y técnicas de compresión de datos.
La implementación HOLAP es un híbrido que combina ROLAP Y MOLAP tratando de obtener los beneficios de cada una de sus implementaciones.

2.6 Diseño conceptual de un Data Warehouse

El diseño conceptual tiene por objetivo la construcción de una descripción abstracta y completa del problema. Comienza con el análisis de requerimientos de los usuarios y de reglas de negocio, a partir del cual se realiza la construcción de un modelo conceptual. En una primera fase se seleccionan los objetos relevantes para la toma de decisiones, y se especifica su utilización como dimensiones o medidas.

Existen dos enfoques diferentes para el diseño conceptual, uno basado en requerimientos y otro basado en los datos fuente. En el primero se analizan los requerimientos de los usuarios y se identifican en ellos los hechos, dimensiones y medidas relevantes. A partir de esto se modela la realidad como un conjunto de cubos multidimensionales.

En el segundo enfoque se construyen cubos multidimensionales transformando un esquema conceptual de los datos fuente. Se comienza por identificar en el esquema fuente los posibles hechos relevantes para la toma de decisiones. A partir de los hechos identificados navegan por las entidades y relaciones construyendo las jerarquías de las dimensiones [5].

2.7 Diseño lógico de un Data Warehouse

A partir del diseño conceptual, se genera un esquema lógico relacional o multidimensional que satisfaga tanto los requerimientos funcionales como los no funcionales. Estos últimos incluyen los requerimientos de rendimiento en la realización de consultas complejas para el análisis de la información, así como también las estrategias de almacenamiento de los datos. Las estructuras básicas en el diseño lógico de un DW son las tablas de hechos (Fact tables) y las tablas de dimensiones (Dimension tables). Las tablas de hechos son estructuras centrales dentro de un esquema lógico multidimensional que contienen los valores de las medidas de negocio que interesan analizar. Las tablas de dimensiones se vinculan con las Tablas de hechos y determinan parámetros de los cuales dependen las medidas registradas en ellas [6].

Hay varios esquemas posibles para estructurar dichas tablas en un modelo lógico. A continuación se describen algunos de ellos.

Star schema

Su estructura contiene una tabla de hechos con los datos de un evento de negocio y un conjunto de tablas de dimensiones organizadas a su alrededor. En estas últimas se colapsan (o desnormalizan) las jerarquías de las dimensiones del modelo original, de forma tal que cada tabla de dimensión puede contener una multiplicidad de jerarquías independientes. La tabla de hechos se vincula con las dimensiones en una relación N-1, en donde la clave primaria de la tabla de hechos es la concatenación de las claves primarias de todas las tablas de dimensiones que la rodean [6]. En la Figura 2-7 se muestra un ejemplo de Star schema.

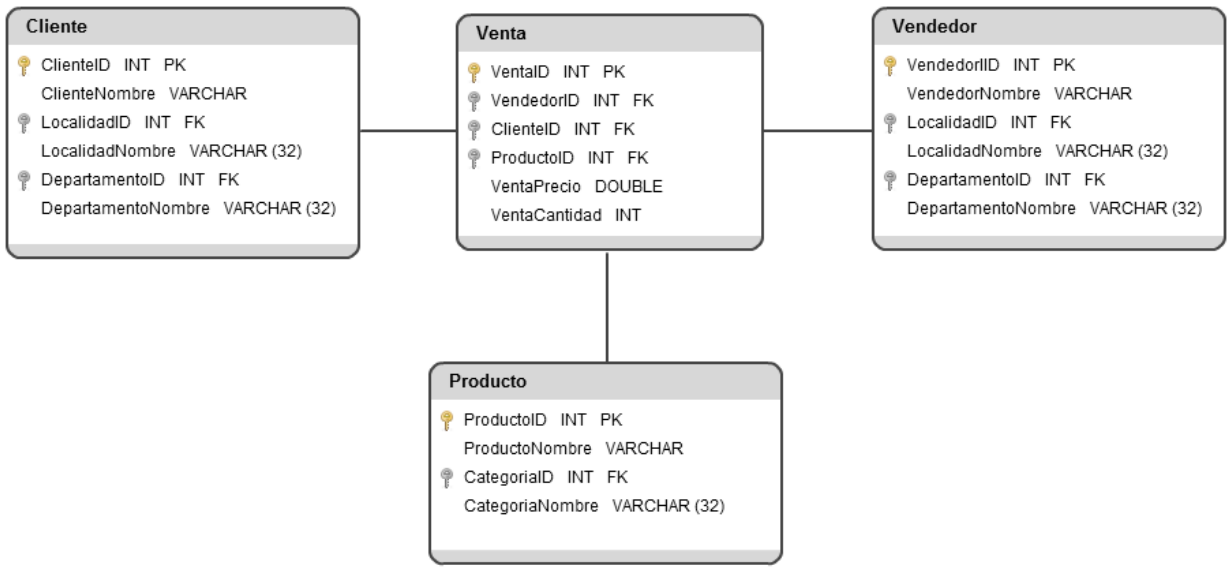


Figura 2-7: Star Schema

Snowflake schema

A diferencia de lo que ocurre en un *Star schema*, en un *Snowflake schema* se normalizan las jerarquías de cada una de sus dimensiones, quedando estas últimas estructuradas en más de una tabla [6]. La Figura 2-8 muestra un ejemplo de Snowflake schema.

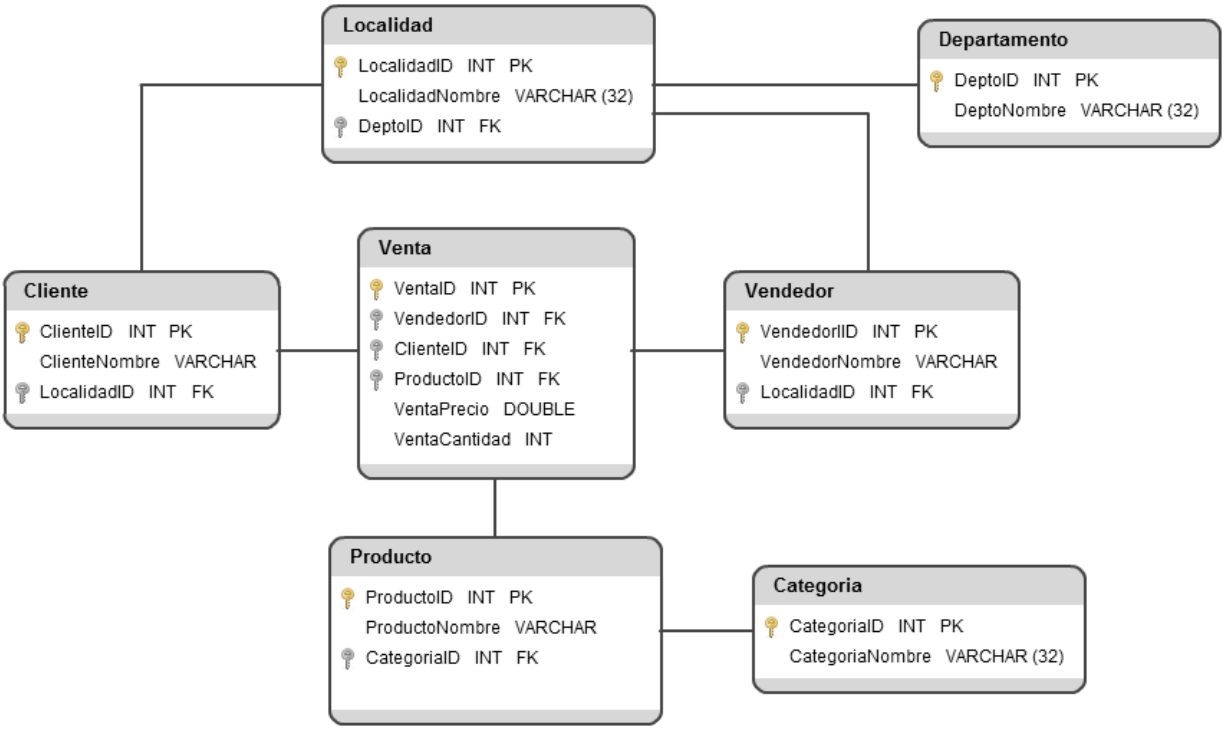


Figura 2-8: Snowflake Schema

Star cluster schema

Es una combinación de *Star chema* con *Snowflake schema*. Selectivamente se normalizan los fragmentos de jerarquías compartidos entre diferentes dimensiones. El resto de las jerarquías se mantienen desnormalizadas [7]. Un ejemplo de
Star cluster schema muestra en la Figura 2-9.

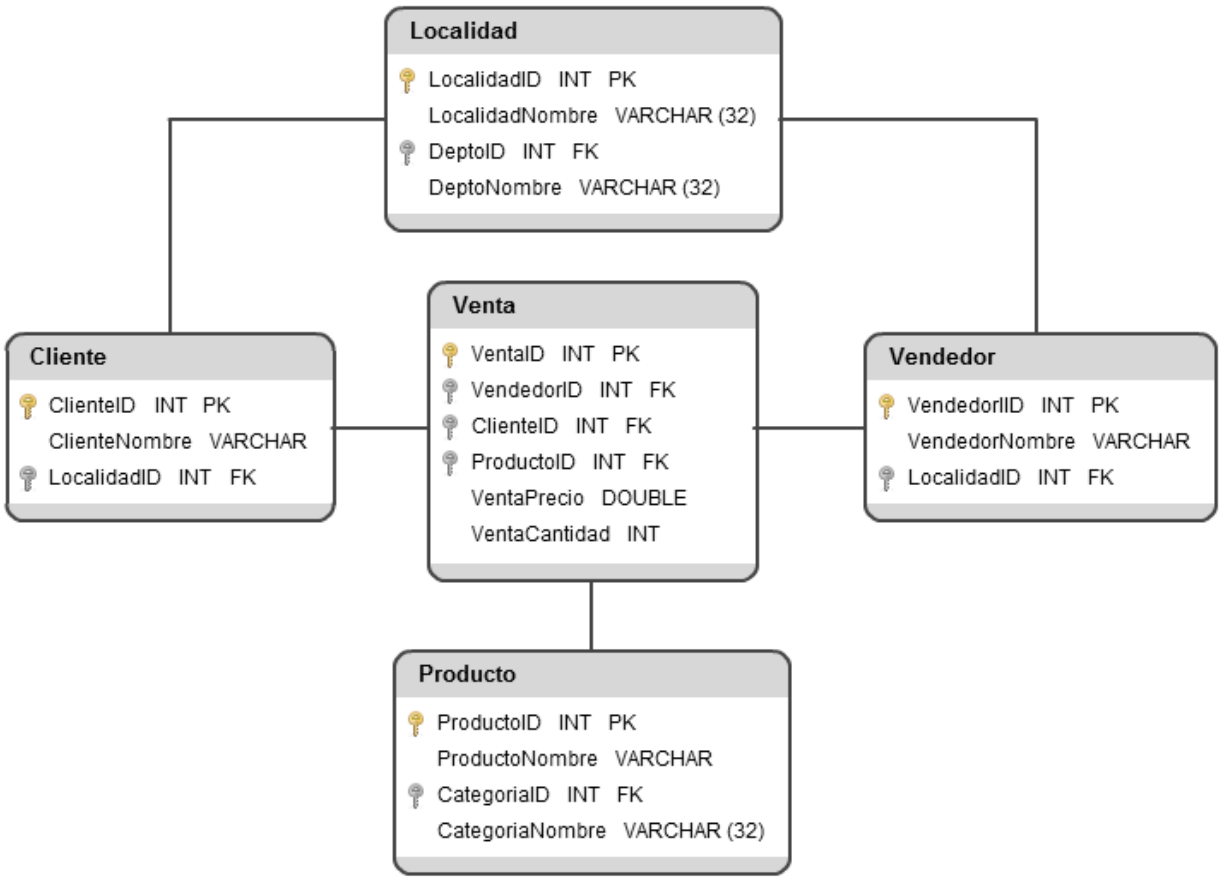


Figura 2-9: Star cluster schema

2.8 ETL

El acrónimo ETL (Extract, Transform and Load) refiere al proceso de Extracción, Transformación y Carga de datos desde múltiples fuentes hacia bases de datos, DW o cualquier otro Sistema de Información. Su principal objetivo es mantener cargado el sistema con los datos correspondientes, los que pueden ser limpiados y reformateados durante el proceso. La primera etapa consiste en extraer los datos de las fuentes de origen, siendo muy común que dichas fuentes sean diversas. La etapa de transformación aplica una serie de reglas de negocio sobre los datos extraídos para convertirlos en los datos que serán cargados en los sistemas destino. Esto último es lo que conforma la etapa de carga.

El ciclo de vida de un ETL consiste en tres grandes etapas: diseño, carga inicial y actualización. En la primera etapa se diseñan las estructuras de datos, así como también los procesos de carga y actualización. Se realiza la especificación del esquema del DW y se definen las entidades encargadas de la extracción, limpieza e integración de los datos.

En la segunda etapa se genera el contenido inicial con que se alimenta el DW. Consta de cuatro actividades. La primera consiste en la Preparación, en donde se realiza la extracción, limpieza y almacenamiento de datos de cada fuente. Luego se realiza la Integración de datos provenientes de fuentes heterogéneas. La tercera actividad es el Agrupamiento, que consiste en la generación, a partir de las vistas base, de otras vistas agrupadas y resumidas. La última actividad es la Adaptación, en donde se generan vistas de usuario que definen los DM. La etapa de actualización busca resolver el problema de cómo impactar los cambios ocurridos en los sistemas fuente en el DW. Tiene un flujo de datos similar a la carga y se ejecuta con una frecuencia planificada.

Operaciones

Dentro de la estructura general de los procesos ETL existen diversas operaciones para la manipulación de los datos. Entre las operaciones básicas se encuentran la identificación de las respectivas fuentes de datos, la extracción de datos de dichas fuentes, el filtrado y la transformación de datos, y la integración de diversas estructuras origen en una estructura destino. Además de las operaciones básicas existen otras relacionadas con la semántica de las aplicaciones, como la Historización, que agrega una marca de tiempo a los datos, o la realización de determinados cálculos encargados de la generación de nuevas variables derivadas de otras existentes. Dentro de este grupo de operaciones se puede considerar a la Limpieza de Datos, que está presente en la mayoría de los procesos de migración de datos, y tiene como objetivo que los datos que se obtienen como resultado de la ejecución del proceso tengan una mejor calidad que los datos de entrada. Más adelante, dentro de este mismo capítulo, se trata este aspecto con más profundidad [8].

Flujos de control y de datos

Durante la ejecución de un proceso ETL las operaciones se realizan en un cierto orden. La sucesión de dichas operaciones determinan el flujo de control del proceso ETL. Existe también un flujo de datos que determina la conexión entre las diferentes operaciones y no necesariamente implica un orden de ejecución. La entrada de una operación es la salida de otras.

Herramientas ETL

Las herramientas ETL son ambientes especializados que permiten la definición y manipulación de objetos en aplicaciones de intercambio de datos. Brindan funcionalidades que, aplicando transformaciones, facilitan el proceso de migración de datos entre diferentes sistemas, y son, generalmente, las primeras en hacer contacto con las fuentes de datos.

2.9 Limpieza de datos

Los sistemas de DW suelen cargar grandes cantidades de datos provenientes de fuentes heterogéneas, por lo que la probabilidad de que contengan datos inconsistentes es alta. Tratándose de sistemas de apoyo a la toma de decisiones, la corrección de los datos que ofician de materia prima cobra vital importancia. Es por esto que se suele invertir buena parte del tiempo del desarrollo de DW en el proceso de limpieza de datos.

La limpieza de datos refiere al acto de detección y corrección de errores presentes en las fuentes de datos de un sistema. El proceso de limpieza identifica inconsistencias en los datos y permite su modificación o eliminación, con el fin de mejorar la calidad de los datos que son cargados. Para efectuar el proceso de limpieza se transita por las siguientes etapas: Análisis de datos, Definición de los procesos de limpieza, Ejecución y Post-proceso [9].

Análisis de datos
Durante esta etapa se trata de identificar qué tipo de errores e inconsistencias presentan las fuentes de datos. Con este fin, se realiza una inspección sobre dichas fuentes para detectar datos corruptos, incoherentes, erróneos, duplicados e incompletos. Luego de la ejecución de esta etapa se tiene un panorama inicial del tipo de errores que presentan los datos y, en virtud de esto, se decide el tipo de estrategia que se utilizará para realizar la limpieza.

Definición de los procesos de limpieza
En esta etapa se definen las tareas de limpieza a ejecutar y el lugar donde se aplicarán. También se define si estás tareas pueden ser automatizadas o se deben ejecutar manualmente.

Ejecución
Se procede a la ejecución de las tareas definidas en el paso anterior.

Post-proceso
En esta última etapa los datos que no pueden ser corregidos automáticamente durante la etapa de ejecución se corrigen, en la medida de lo posible, de forma manual.

2.10 Dimensiones de Calidad de Datos

La calidad de los datos es caracterizada a través de múltiples dimensiones que sirven para evaluar y calificar los datos que componen un Sistema de Información. Cada una de estas

dimensiones refleja un aspecto particular de la Calidad de Datos [10], y para cada una de ellas, existen diferentes factores que representan aspectos de calidad particulares de las mismas. A continuación se mencionarán algunas dimensiones, y la elección se realiza en función de su relevancia dentro del marco del proyecto.

Completitud

La completitud es una medida que permite determinar si un sistema contiene toda la información de interés, si representa todos los objetos de la realidad que modela, o qué porción de la realidad está representada. Si se cuenta con todos los datos que describen los objetos de interés, y qué tantos valores son nulos [17].

Exactitud

Intuitivamente, puede verse a la exactitud como una medida que indica qué tan libre de errores están los datos. Esta dimensión mide qué tan correcta y precisa es la relación entre los objetos del mundo real y las entidades que los representan en un Sistema de Información. Existen diferentes factores para la Exactitud, como la Correctidud Semántica, la Correctitud Sintáctica y la Precisión.

La Correctitud Semántica mide qué tan correcta es la correspondencia entre los estados del mundo real y su representación en el Sistema de Información.

La Correctitud Sintáctica mide qué tan correcta es la correspondencia entre los valores del Sistema de Información y los valores válidos de un dominio, sin importar si son valores reales. Por otro lado, la Precisión mide el nivel de detalle de los datos del Sistema de Información [17].

3 Arquitectura del Sistema de Data Warehouse

En este capítulo se describe la arquitectura general de la solución y se detallan sus componentes. Además, se explican las decisiones de diseño en función de los requerimientos no funcionales.

3.1.1 Requerimientos no funcionales

Se requiere que el DW se desarrolle utilizando *Microsoft Sql Server, Standard Edition*, a través de sus herramientas *Integration Services*, para procesos ETL, y *Analysis Services* para la creación de bases de datos multidimensionales. Además, es necesario utilizar el producto de Business Intelligence *O3,* de la empresa *Ideasoft*, como complemento a *Analysis Services*, para el diseño de los cubos y la presentación de la información a los usuarios. Por otro lado, teniendo en cuenta que en el MIDES se está estudiando la posibilidad de migración de *Microsoft Sql Server* a otro motor de base de datos, también se requiere probar alguna herramienta open source como alternativa a *Integration Services*.

3.1.2 Decisiones de diseño

La Figura 3-1 presenta la arquitectura general de la solución, en donde se pueden observar sus diferentes componentes. Dichos componentes se describen a continuación.

Fuentes de datos
Este componente contiene las diversas fuentes de datos utilizadas para alimentar el DW. Para este trabajo, dichas fuentes consisten en bases de datos relacionales *Oracle* y *SQLServer*, y hojas de cálculo de Microsoft Excel.

ODS
Algunas fuentes de datos suelen variar su estructura en diferentes instancias de actualización del DW. Por ejemplo, las hojas de cálculo de Excel, aun manteniendo los datos requeridos, pueden cambiar el nombre y tipo de dato de sus campos, así como también el orden de las columnas que los contienen. La cantidad de columnas de estos archivos también suele variar, agregándose o eliminándose variables superfluas al interés del análisis. Previendo, además, la incorporación de otras fuentes que pueden presentar problemas similares como pueden ser archivos de texto, se decidió crear un ODS. Éste consiste en una base de datos intermediaria entre los datos fuente puros, sin procesar, y los datos del DW. En dicho componente se integran los datos de las distintas fuentes y se

realizan las primeras transformaciones para limpiar y preparar los datos que luego serán cargados al DW. En general, se realiza limpieza que no se puede automatizar.
Se puede decir que el ODS funciona como una interfaz entre las fuentes de datos y el DW, en donde quedan bien definidos los datos necesarios para las futuras actualizaciones del mismo. Esto permite que los procesos de carga más complejos, que son los que van desde el ODS al DW, sean más claros y estables.

ETL
Este componente contiene el conjunto de tareas necesarias para realizar el proceso de carga y actualización del DW. Dicho proceso implica la extracción, filtrado, trasformación, integración, limpieza y carga de datos al sistema.

Data Warehouse
Los datos del ODS, transformados por los procesos de carga ETL, son almacenados en una base de datos relacional no normalizada, en tablas de hechos y dimensiones. Esta base de datos integra la información de todos los programas sociales por medio de las dimensiones compartidas, constituyendo la base del DW corporativo. La utilización de esta base permite independizarse de la herramienta OLAP que se elija para el análisis. De los requerimientos no funcionales se desprende que la solución debe ser flexible en este aspecto.
Como desventaja de la utilización de esta base se observa que según la herramienta de análisis que se utilice, puede existir una redundancia de datos, una vez en la base relacional y otra en un modelo multidimensional. En este trabajo los prototipos realizados no consultan directamente la base del DW sino que la misma es utilizada para generar las bases multidimensionales de los Data Marts.

Data Marts
Este componente contiene las bases de datos multidimensionales orientadas al análisis de la información de cierta área del negocio y orientada a un grupo específico de usuarios.
Con los datos extraídos de la base del DW se construyen los cubos multidimensionales desde donde se obtiene la información para el análisis OLAP.

Presentación
Esta es la capa destinada a la interacción del usuario con el DW. Para la presentación se utilizó Microsoft Excel para consultar los cubos desde *Analysis Services* y *O3 Portal*, herramienta web proporcionada por *Ideasoft*, para los cubos generados con *O3*.

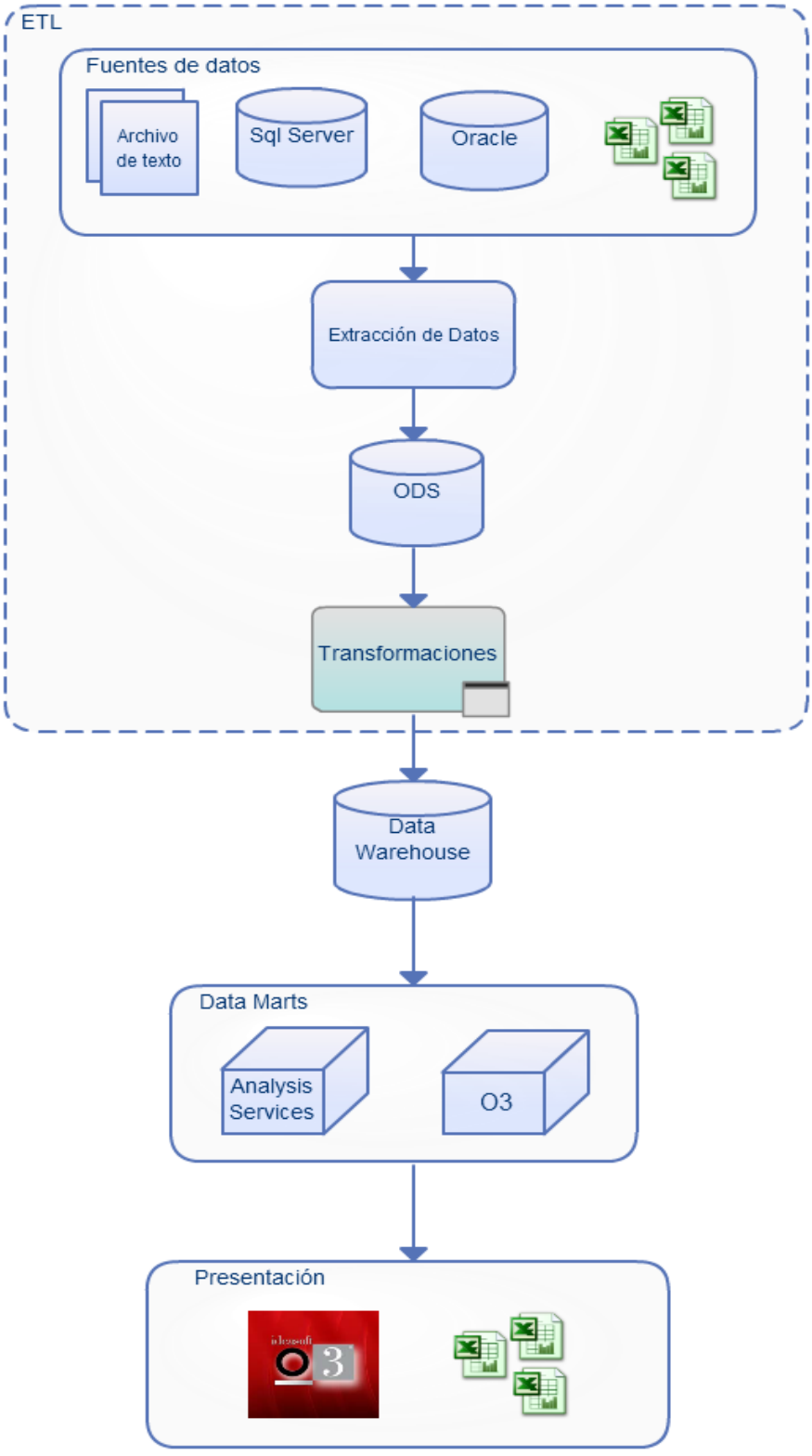


Figura 3-1: Arquitectura general

4 Análisis y diseño

Este capítulo contiene el análisis, el diseño conceptual y el diseño lógico del DW para los dos programas sociales seleccionados. Para cada uno de ellos se realiza una breve descripción introductoria. Luego se describen los requerimientos y las características de las fuentes de datos que alimentan al DW. Finalmente se presentan el modelado conceptual multidimensional y el diseño lógico de la solución.
Para el diseño conceptual se utilizó el enfoque basado en requerimientos (ver sección 2.6). Mientras que para la representación de los esquemas multidimensionales se utilizó el modelo *Conceptual MultiDimensional Model* (CMDM) [11].

4.1 Uruguay Integra

El programa *Uruguay Integra* pertenece al conjunto de programas, implementados por el MIDES, orientados a la seguridad y protección social. Es un programa socio-educativo que apunta al desarrollo de habilidades que promueven la inclusión e integración social. Se trata de un espacio pensado para compartir, aprender, recrearse y capacitarse. Su objetivo es brindar oportunidades de desarrollo personal e integración ciudadana a sectores de la población en situación de vulnerabilidad social, buscando que sus participantes desarrollen las siguientes habilidades:

- reconocimiento y promoción de derechos.
- abordaje y seguimiento de situaciones-problemas identificados.
- revalorización del trabajo.
- desarrollo del lenguaje y del razonamiento.
- intercambio e integración con la comunidad.

Este programa es dirigido a hombres y mujeres mayores de 18 años, sin importar su nivel educativo [12]. Se implementa a partir de talleres coordinados por OSC (Organización de la Sociedad Civil) que tienen convenio con el MIDES, a través de Equipos Técnicos responsables de la implementación y seguimiento del programa.
Los talleres, a través de una actividad específica o central, abordan dos tipos de actividades, las socio-educativas o de promoción social y las de vínculo o integración social. Tienen un cupo mínimo de 15 y un máximo de 20 participantes, y una carga horaria mínima de 6 horas semanales, durante 8 meses.
Las personas inscriptas que cumplan los requisitos para participar en el programa son distribuidas en grupos. Cada uno de estos grupos es dirigido por varios técnicos de una OSC y un técnico del MIDES. Una OSC dirige varios grupos, y cada uno de sus técnicos puede

trabajar en varios grupos. Existen grupos en varias localidades de todos los departamentos del país.

A continuación se definen conceptos del programa que son utilizados a lo largo del documento.

Inscripto
Persona que se inscribió en el programa y fue habilitada para realizar los talleres.

Beneficiario
Persona que se inscribió en el programa y asistió, como mínimo, a un 75% de los talleres realizados.

OSC
Organización de la Sociedad Civil. Son organismos que trabajan en conjunto con técnicos del MIDES para llevar a cabo la implementación del programa.

Edición
Versión del programa establecida en un contexto temporal. Las ediciones del programa *Uruguay Integra* son anuales.

Modalidad
La implementación del programa se divide en diferentes modalidades elegidas de forma arbitraria. Éstas pueden indicar, por ejemplo, el contexto geográfico en el cual se lleva a cabo dicha implementación.

Grupo
Conjunto de personas designadas para participar de un taller. Las designaciones obedecen generalmente a criterios geográficos.

4.1.1 Requerimientos de análisis de datos

Según los requerimientos planteados por los usuarios se distinguen tres formas de análisis para los datos:

- Según los recursos humanos asociados al programa.
- Según el contexto, enfocado al estudio de las características de la población de participantes.
- Análisis de producto, en base a las características del programa y su ejecución.

A continuación se detallan los requerimientos de cada uno de los enfoques mencionados:

Análisis de los datos en base a los recursos humanos

UIR-1 Grupos
Se quiere conocer la cantidad de grupos creados por edición, zona geográfica y modalidad de programa.

UIR-2 RRHH del programa - Técnicos
Se quiere saber la cantidad de técnicos, de cada OSC, que trabajaron en cada edición del programa, discriminados por grupo, zona y modalidad.

UIR-3 RRHH del programa - Funcionarios
Se quiere saber la cantidad de funcionarios del MIDES que trabajaron en cada edición del programa, discriminados por grupo, modalidad y zona geográfica, y también con qué OSC lo hicieron.

UIR-4 RRHH del programa – Beneficiarios
Se quiere saber la cantidad de beneficiarios que integraron los talleres brindados por grupo, edición, modalidad, zona geográfica y OSC.

Análisis de los datos en base al contexto

UIC-1 Distribución de la población de inscriptos al programa
Se desea obtener la cantidad de personas inscriptas al programa en cada edición por: sexo, diferentes tramos de edad, nivel educativo alcanzado, situación laboral, OSC, grupo, zona geográfica y modalidad de programa. Este análisis se debe poder realizar teniendo en cuenta las personas inscriptas, así como también a los beneficiarios.

UIC-2 Menores por hogar
Se desea obtener la cantidad de menores de 18 años en los hogares de los beneficiarios, estos últimos discriminados por sexo, edad, nivel educativo, situación laboral, grupo, modalidad de programa y zona geográfica.

UIC-3 Relación entre personas al comienzo y fin del programa
Se desea poder ver la relación entre la cantidad de inscriptos y los beneficiarios, discriminados por sexo, edad, nivel educativo, situación laboral, grupo, modalidad de programa y zona geográfica. Se busca obtener los valores como porcentaje.

Análisis de los datos en base al producto

UIP-1 Cantidad de cupos
Se desea saber la cantidad de cupos generados para el programa en cada edición, discriminados también por modalidad y zona geográfica.

UIP-2 Relación inscriptos - cupos
Se desea saber la cantidad de inscriptos por cupos generados. O sea, el cociente entre el total de inscriptos y la cantidad de cupos máximos generados. Interesa realizar el análisis discriminando por grupo, modalidad y zona geográfica.

UIP-3 Abandonos
Se desea ver la cantidad de abandonos desagregados por motivo, grupo, edición, modalidad y zona geográfica.

4.1.2 Fuentes de datos

Los datos utilizados para alimentar el DW están contenidos en archivos Excel. Si bien éstos suelen variar en cuanto a su cantidad y estructura, en general, existen dos por cada edición del programa. Uno de ellos contiene una fila por cada participante del programa. En cada una de dichas filas se especifica la edición del programa, la modalidad, el departamento, la localidad, grupo al que perteneció el participante, el nombre de la OSC que trabajó con dicho grupo, el funcionario del MIDES que trabajó en ese grupo, datos del participante y de su hogar, un indicador de si abandonó el programa y, si fue así, el motivo de su abandono. La Tabla 4-1 muestra un esquema con los datos de esta planilla utilizados para la carga del DW.
La otra planilla contiene información de los técnicos de las OSC que trabajaron en el programa. Cada fila contiene los datos de un técnico, incluyendo un grupo en el que haya trabajado, por lo que la cantidad de filas que le corresponden a cada técnico es equivalente a la cantidad de grupos en los cuales trabajó. La Tabla 4-2 muestra un esquema con los datos de esta planilla utilizados para la carga del DW.

Id	Osc	Cedula	NacDia	NacMes	NacAño	Sexo	Depto	Localidad	Abandono	Mot_Aba	Sit_Lab	Niv_Edu
RU076	AMRU	38.960.149	4	2	1.965	1	18	Peralta	3	6	98	0
	3	42.535.376	6	3	1.972	2	16	Ciudad del pla	1	10	4	2
1602.0		19.413.856	24	12	1.961	2	16	Ciudad de l Pl	1	1	3	2
1601.0	3	33.064.396	27	12	1.971	2	16		2			6

Tabla 4-1: Esquema de planilla de participantes de UI.

Identificador	Nombre_OSC	CI	Nac_dia	Nac_mes	Nac_año	Sexo	Depto	Localidad
RU076	Asociación Mujeres Rurales Uruguayas	31730315	18	8	1.976	1	18	
UA012-1	Umbrales	35713779	27	3	1.977	2	17	Mercedes
UA012-2	Umbrales	37679618	12	5	1.966	2	17	Mercedes
RU113	KÖLPING	948418	5	4	1.941	2	6	Durazno
768	Asociación Civil Umbrales	18786252	24	7	1.961	1	Artigas	Artigas

Tabla 4-2: Esquema de planilla de técnicos de UI.

4.1.3 Diseño conceptual

Aquí se presenta el modelo conceptual, el cual permite observar el alcance del trabajo para el programa *Uruguay Integra*. Se indican qué datos son interesantes y cómo se relacionan entre sí. Como primer paso se identifican dimensiones y medidas, luego se muestra cómo éstas se relacionan para satisfacer los requerimientos. Además, se exhiben las relaciones dimensionales y el cuadro de roll-up.

4.1.3.1 Dimensiones

A continuación se describen las dimensiones del programa Uruguay Integra.

- **OSC**

  La dimensión *OSC* representa las organizaciones de la sociedad civil. En el programa UI estas organizaciones son las encargadas de trabajar con los grupos. Es utilizada para el análisis de los recursos humanos con los que cuenta el programa (trabajadores y participantes) según la OSC con las cuales trabajaron (Figura 4-1).

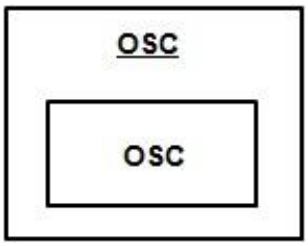


Figura 4-1: Dimensión OSC

- **Grupo**

  La dimensión *Grupo* representa los grupos del programa Uruguay Integra. Esta dimensión permite agregar los grupos según la zona geográfica, la edición y la modalidad. Para esto, la dimensión está estructurada en varias jerarquías de agregación. Respecto a la jerarquía geográfica, los grupos se agregan en localidades, y las localidades en departamentos. Las otras dos jerarquías agregan los grupos por edición y por modalidad (Figura 4-2).

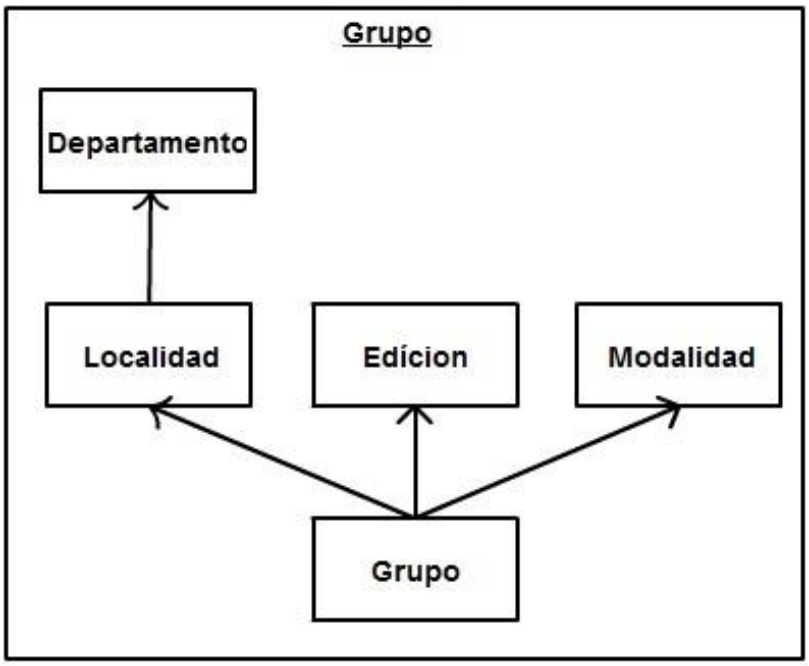


Figura 4-2: Dimensión Grupo

- **Sexo**

  La dimensión *Sexo* representa el género de la persona (Figura 4-3).

Figura 4-3: Dimensión Sexo

- **Edad**

  La dimensión *Edad* es utilizada para el análisis de las características de los participantes del programa según las edades de los mismos. Esta dimensión se estructura en una jerarquía de dos niveles (Figura 4-4), en donde las edades se agrupan en diferentes rangos, los cuales fueron definidos con los usuarios de la siguiente forma:

  - Menor de 18
  - De 18 a 25 años
  - De 26 a 30 años
  - De 31 a 35 años
  - De 36 a 40 años
  - De 41 a 50 años
  - De 51 a 60 años
  - De 61 a 70 años
  - De 71 a 80 años
  - De 81 a 90 años

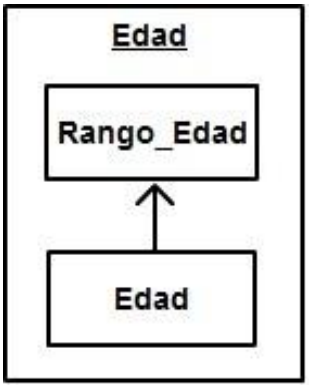


Figura 4-4: Dimensión Edad

- **Nivel Educativo**

  La dimensión *Nivel Educativo* representa el máximo nivel de educación alcanzado por una persona (Figura 4-5). Esta dimensión es utilizada para el análisis de las características de los participantes del programa. A continuación se listan los valores posibles de la dimensión:

  - Sin instrucción
  - Preescolar
  - Primaria incompleta
  - Primaria completa
  - Ciclo Básico incompleto
  - Ciclo Básico completo
  - Bachillerato incompleto
  - Bachillerato completo
  - Enseñanza Técnica incompleta
  - Enseñanza Técnica completa
  - Enseñanza Militar o Policial incompleta
  - Enseñanza Miliar o Policial completa
  - Terciaria no universitaria incompleta
  - Terciaria no universitaria completa
  - Universidad o similar incompleta
  - Universidad o similar completa
  - Sin dato

Figura 4-5: Dimensión Nivel Educativo

- **Situación Laboral**

  La dimensión *Situación Laboral* identifica aspectos de la relación entre el participante y su actividad laboral durante la semana previa a la inscripción al programa (Figura 4-6).

  Al igual que en el caso del nivel educativo, esta dimensión también es utilizada para el análisis de las características de los participantes del programa. A continuación se listan sus valores posibles:

  - Trabajó
  - No trabajó pero si tenía trabajo
  - No trabajó pero buscó
  - Se dedica a las tareas del hogar pero no busca trabajo remunerado
  - Es jubilado pensionista
  - Tiene una discapacidad que le impide trabajar en forma permanente
  - No trabaja y no busca trabajo
  - Sin Dato

Figura 4-6: Dimensión Situación Laboral

- **Motivo**

  El motivo representa la razón por la cual un beneficiario abandona el programa.
  Esta dimensión es utilizada para el análisis de los participantes que abandonaron el programa antes de su culminación (Figura 4-7). Los valores posibles para esta dimensión son los siguientes:

  - Actividades laborales
  - Cuidados de enfermos en el lugar
  - Cuidados de menores en el hogar
  - Cuidado del hogar (seguridad)
  - Estudios
  - Falta de interés
  - Mudanzas
  - Problemas de salud
  - Tareas domésticas
  - Otros motivos
  - No corresponde
  - No abandonó
  - Sin dato

Figura 4-7: Dimensión Motivo

4.1.3.2 Medidas

- **Cantidad de grupos**: Representa la cantidad de los grupos creados en el programa.

- **Cantidad de técnicos**: Representa la cantidad de los técnicos que trabajaron en el programa.

- **Cantidad de funcionarios**: Representa la cantidad de los funcionarios MIDES que trabajaron en el programa.

- **Cantidad de beneficiarios**: Representa la cantidad de los beneficiarios que participaron en el programa.

- **Cantidad de inscriptos**: Representa la cantidad de las personas inscriptas para participar en el programa.

- **Cantidad de menores de 18 años**: Representa la cantidad de las personas menores de 18 años en los hogares de los inscriptos y beneficiarios.

- **Relación beneficiarios inscriptos**: Medida calculada que representa el cociente entre las medidas *Cantidad de beneficiarios* y *Cantidad de inscriptos*.

- **Cantidad máxima de cupos**: Representa la cantidad máxima de cupos creados por cada grupo.

- **Cantidad mínima de cupos**: Representa la cantidad mínima de cupos creados por cada grupo.

- **Relación inscriptos grupos**: Medida calculada que representa el cociente entre las medidas *Cantidad de inscriptos* y *Cantidad de cupos máximo*.

- **Cantidad de abandonos**: Representa la cantidad de participantes que abandonaron el programa.

4.1.3.3 Relación entre dimensiones, medidas y requerimientos

Esta sección muestra cómo se utilizan las diferentes dimensiones y medidas para satisfacer los requerimientos solicitados por los usuarios.

Recursos humanos

UIR-1 Grupos

- **Dimensiones:**
    - ***Grupo:*** Interesa conocer la cantidad de grupos por edición de programa, modalidad y zona geográfica.

- **Medidas**
    - ***Cantidad de grupos.***

UIR-2 RRHH del programa - Técnicos

- **Dimensiones:**
    - ***OSC:*** Interesa conocer la cantidad de técnicos que trabajaron en el programa discriminados por OSC.
    - ***Grupo:*** Interesa conocer la cantidad de técnicos por grupo, edición de programa, modalidad y zona geográfica.

- **Medidas**
    - ***Cantidad de técnicos.***

UIR-3 RRHH del programa - Funcionarios

- **Dimensiones:**
    - ***OSC:*** Interesa conocer con cuál OSC trabajó cada funcionario.
    - ***Grupo:*** Interesa conocer la cantidad de funcionarios por grupo, edición y modalidad de programa y zona geográfica.

- **Medidas**
    - ***Cantidad de funcionarios.***

UIR-4 RRHH del programa – Beneficiarios

- **Dimensiones:**
    - ***OSC:*** Interesa conocer a cuál OSC corresponde el taller del cual participó cada beneficiario.
    - ***Grupo:*** Interesa conocer la cantidad de beneficiarios por grupo, edición y modalidad de programa y zona geográfica.

- **Medidas**
    - ***Cantidad de beneficiarios.***

Contexto

UIC-1 Distribución de la población de inscriptos al programa

- **Dimensiones:**
  - ***Sexo:*** Interesa clasificar a los inscriptos y beneficiarios por sexo.
  - ***Edad→Rango_Edad:*** Interesa clasificar a los inscriptos y beneficiarios según diferentes tramos de edad.
  - ***Nivel Educativo:*** Interesa clasificar a los inscriptos y beneficiarios según su nivel educativo.
  - ***Situación Laboral:*** Interesa clasificar a los inscriptos y beneficiarios según su condición de actividad laboral.
  - ***Grupo:*** Interesa clasificar a los inscriptos y beneficiarios de acuerdo a la edición, modalidad y grupo en el cual participaron, y al departamento y localidad a la cual pertenecen.
  - ***OSC:*** Interesa clasificar a los inscriptos y beneficiarios según la OSC que trabajó con su grupo.

- **Medidas**
  - ***Cantidad de Inscriptos.***
  - ***Cantidad de Beneficiarios.***

UIC-2 Menores por hogar

- **Dimensiones:** son las mismas que para el requerimiento **UIC-1.**

- **Medidas**
  - ***Cantidad de menores de 18 años.***

UIC-3 Relación entre personas al comienzo y fin del programa

- **Dimensiones:** son las mismas que para el requerimiento **UIC-1.**

- **Medidas**
  - ***Relación beneficiarios inscriptos.***

Producto

UIP-1 Cantidad de cupos

- **Dimensiones:**
  - ***Grupo:*** Interesa conocer la cantidad de cupos por edición, modalidad y zona geográfica.

- **Medidas**
  - ***Cantidad máxima de cupos.***
  - ***Cantidad mínima de cupos***.

UIP-2 Relación inscriptos – cupos

- **Dimensiones:**
  - ***Grupo:*** Interesa conocer la relación entre inscriptos y cupos por edición, modalidad y zona geográfica.

- **Medidas**
  - ***Relación inscriptos grupos***.

UIP-3 Abandonos

- **Dimensiones:**
  - ***Grupo:*** Interesa conocer la cantidad de abandonos por grupo, edición, modalidad y zona geográfica.
  - ***Motivo:*** Interesa clasificar los abandonos en función del motivo.

- **Medidas**
  - ***Cantidad de abandonos.***

4.1.3.4 Relaciones dimensionales

Se definieron las relaciones dimensionales *Recursos*, *Contexto* y *Producto*, de acuerdo a las perspectivas de análisis definidas junto con el usuario.

- **Recursos**

  En la Figura 4-8 se muestra la relación dimensional *Recursos*, la cual modela la información acerca de los recursos humanos utilizados para la ejecución del programa. Sus dimensiones son *OSC* y *Grupo*, y las medidas son *Cantidad de Grupos*, *Cantidad de técnicos*, *Cantidad de Funcionarios* y *Cantidad de Beneficiarios*.

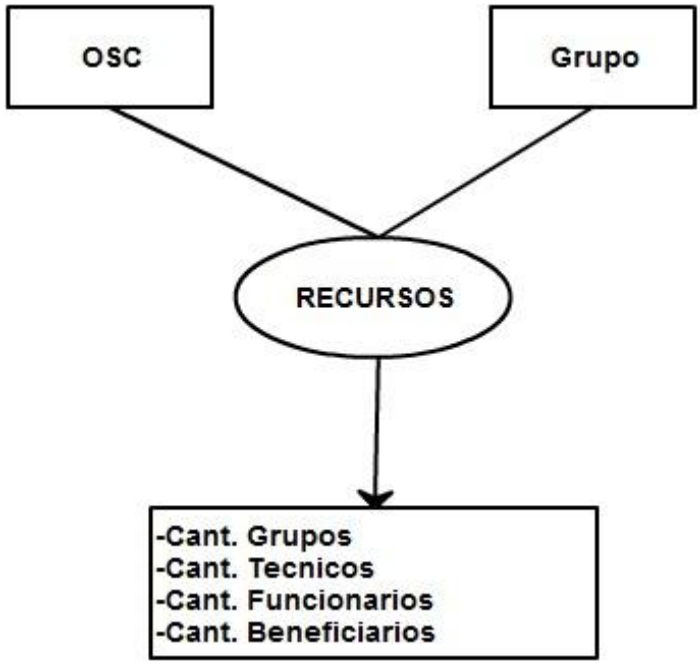


Figura 4-8: Relación Dimensional Recursos

- **Contexto**

  La Figura 4-9 muestra la relación dimensional *Contexto*, la cual modela la información sobre el contexto de la población que participa en el programa. Sus dimensiones son *OSC, Grupo, Sexo, Edad, Nivel Educativo y Situación Laboral*, y las medidas son *Cantidad de Inscriptos*, *Cantidad de Beneficiarios, Relación Beneficiarios Inscriptos* y *Cantidad de menores de 18 años*.

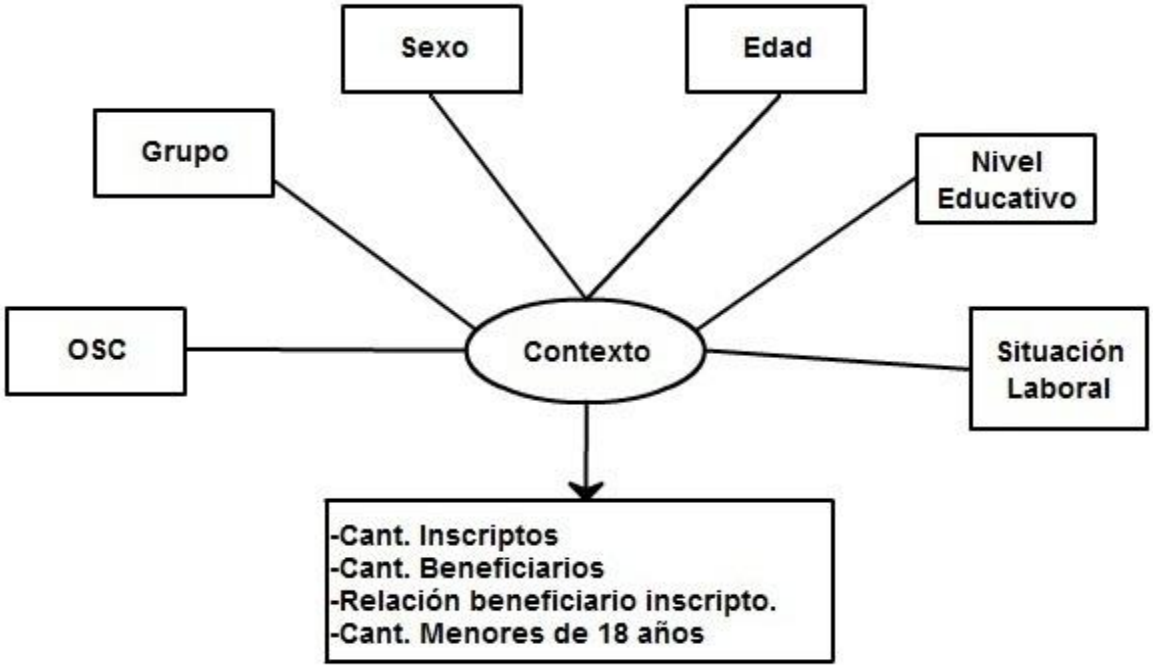


Figura 4-9: Relación Dimensional Contexto

- **Producto**

  La Figura 4-10 muestra la relación dimensional *Producto*, la cual modela la información sobre el resultado del programa. Sus dimensiones son *Motivo* y *Grupo*, y las medidas son *Cantidad máxima de cupos*, *Cantidad mínima de cupos*, *Cantidad de abandonos* y *Relación inscriptos cupos*.

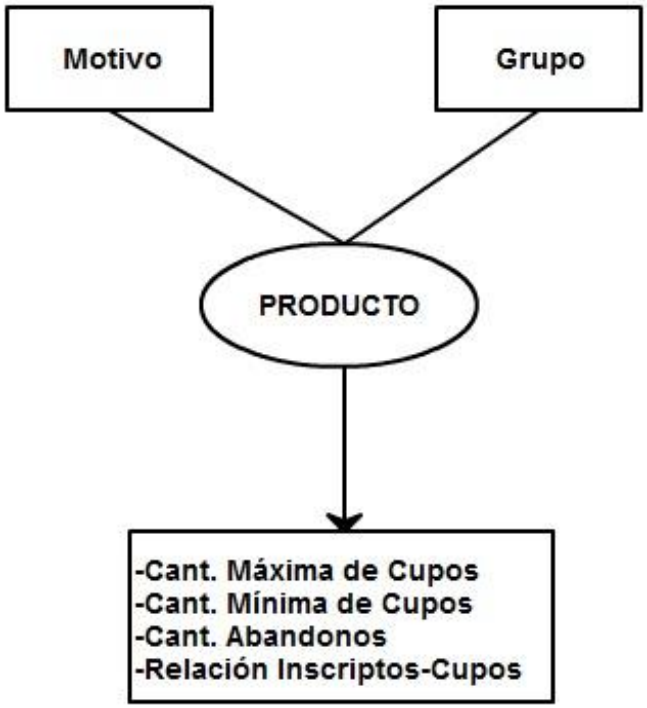


Figura 4-10: Relación Dimensional Producto

4.1.3.5 Cuadro de Roll Up

Esta sección contiene los Cuadros de Roll Up, los cuales muestran en forma esquemática cómo se calculan las medidas al realizar los agrupamientos por las distintas jerarquías de las dimensiones.

- **Recursos**

		Cant. Grupos	Cant. Técnicos	Cant. Funcionarios	Cant. Beneficiarios
Grupo	**Grupo→Modalidad**	+	+[1]	+	+
	Grupo→Edición	+	+[1]	+	+
	Grupo→Localidad	+	+[1]	+	+
	Grupo→Localidad→Departamento	+	+[1]	+	+

[1] Existe el problema de doble conteo pues un técnico puede trabajar en más de un grupo.

- **Contexto**

		Cant. Inscriptos	Cant. Beneficiarios	Relación beneficario inscripto	Cant. Menores de 18 años
Grupo	**Grupo→Modalidad**	+	+	2	+
	Grupo→Edición	+	+	2	+
	Grupo→Localidad	+	+	2	+
	Grupo→Localidad→Departamento	+	+	2	+
Edad	**Edad→Rango_edad**	+	+	2	+

[2] Se calcula después de agregar.

- **Producto**

		Cant. Máxima de Ccupos	Cant. Mínima de cupos	Cant. Abandonos	Relación inscripto cupos
Grupo	**Grupo→Modalidad**	+	+	+	3
	Grupo→Edición	+	+	+	3
	Grupo→Localidad	+	+	+	3
	Grupo→Localidad→Departamento	+	+	+	3

[3] Se calcula después de agregar.

4.1.4 Diseño lógico

Para el diseño lógico de la solución se decidió utilizar un *Star cluster schema* [7].
La decisión de utilizar este modelo se basa en la posibilidad de aprovecharse de los beneficios que brindan los *Star schema* y los *Snowflake schema* combinados. Los primeros permiten un mayor rendimiento al tener una mayor velocidad de acceso a los datos como consecuencia de tener una menor cantidad de tablas de dimensiones. De esta manera, las consultas resultan menos complicadas debido a que realizan uniones y cruzamientos únicamente entre la tabla de hechos y las de dimensiones, sin tener que realizar operaciones de cruzamiento al navegar por las dimensiones.

El *Snowflake schema* tiene la ventaja de evitar redundancia e inconsistencias en los datos de jerarquías compartidas por varias dimensiones, esto implica tener estructuras más complejas debido a la normalización de dimensiones.
Para UI el único fragmento de jerarquía que se normalizó fue el correspondiente a la ubicación geográfica (Localidad->Departamento), de la dimensión grupo.
Si bien la jerarquía correspondiente a la ubicación geográfica de la dimensión *Grupo* no es compartida por ninguna otra dimensión en el diseño de UI, se espera de que sí sea compartida por dimensiones correspondientes al diseño de otros programas sociales, por lo que se considera que esta solución es escalable.

Problema de doble conteo de técnicos

En el modelo dimensional *Recursos* se produce un problema de doble conteo con la medida *Cantidad de técnicos* al agregar por cualquiera de las jerarquías de la dimensión *Grupo*. Esto ocurre debido a que un mismo técnico puede trabajar en más de un grupo, y al sumar la información de esos grupos el técnico es contado dos veces. Este problema se muestra en la Tabla 4-3:

Técnico	Grupo	Localidad	Edición
T1	Grupo1	Loc1	2009
T1	Grupo2	Loc1	2010
T2	Grupo2	Loc2	2010
T3	Grupo3	Loc3	2009
T2	Grupo4	Loc1	2010

Tabla 4-3: Problema de doble conteo. Ejemplar de datos

Dado los valores de origen, la Tabla 4-4-a muestra los valores del resultado de sumar la cantidad de técnicos, de los distintos grupos, al aplicar la operación Roll Up al agrupar por *Localidad*, y la Tabla 4-4-b muestra los valores correctos para esta agregación.

Localidad	Cantidad de Técnicos
Loc1	3
Loc2	1
Loc3	1

4-4-a

Localidad	Cantidad de Técnicos
Loc1	2
Loc2	1
Loc3	1

4-4-b

Tabla 4-4: Aplicación de la operación Roll up

Tabla 4-4-a: Suma de la cantidad de técnicos al agrupar por *Localidad*. **Tabla 4-4-b:** Resultado correcto de la agregación.

Suponiendo que se deseara obtener la cantidad de técnicos que trabajaron en la localidad *Loc1*. Si se sumara la cantidad de técnicos de los distintos grupos de dicha localidad, se obtendría como resultado 3, como se muestra en la tabla 4-4-a. Operando de esta forma, el técnico *T1*, que trabajo con el *Grupo1* y el *Grupo2*, sería contado 2 veces.

Para solucionar este problema se optó por crear otra tabla de hechos llamada *FactRecursosTécnicos*, la cual contiene los identificadores de los técnicos. De esta manera, al agregar la información por la dimensión *Grupo* en cualquiera de sus jerarquías se puede contar la cantidad de técnicos distintos. Cabe destacar que esta tabla solo nos da la información de la medida cantidad de técnicos, la información de las otras medidas se mantienen en la tabla de hechos *FactRecursos*.

Problema de medidas con distinta granularidad

En la relación dimensional *Producto* la medida *Cantidad de abandonos* tiene una granularidad más fina que el resto de las medidas. Esto ocurre debido a que la dimensión *Motivo Abandono* sólo aplica para la medida *Cantidad de abandonos*. Para solucionar este problema se crearon dos tablas de hechos para esta relación dimensional [13]: *FactProductoCuposInscriptos y FactProductoAbandono*. La primera de ellas contempla los requerimientos **UIP-1** y **UIP-2**, y la segunda el requerimiento **UIP-3**.

La Figura 4-11 muestra el diseño lógico de la solución. Se crearon las tablas de hechos *FactRecursos* y *FactContexto* para satisfacer los requerimientos correspondientes al análisis de los recursos humanos (**UIR-1 al UIR-4**) y al análisis en base al contexto (**UIC-1 al UIC-3**) respectivamente.

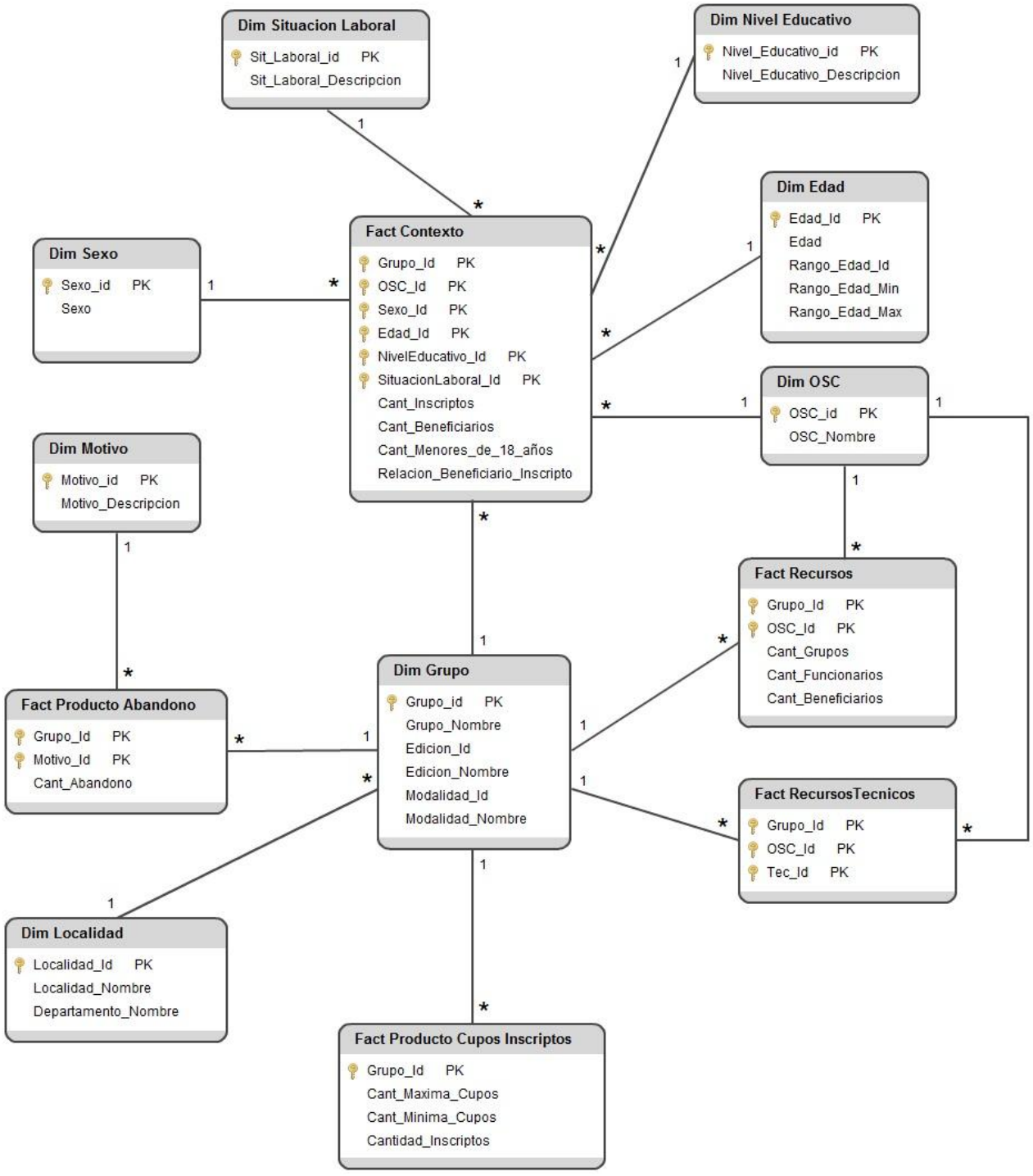


Figura 4-11: Diseño lógico de *Uruguay Integra*

4.2 Tarjeta Uruguay Social

El sistema de compras por la *Tarjeta Uruguay Social* (TUS) funciona desde mayo de 2006 y depende de MIDES, Ministerio de Salud Pública (MSP), Administración de los Servicios de Salud del Estado (ASSE) e Instituto Nacional de Alimentación (INDA). El objetivo principal es permitir que las personas más desprotegidas accedan a los productos de la canasta básica y tengan la posibilidad de seleccionar los productos de acuerdo a sus necesidades y a las características de su núcleo familiar. Otro de sus objetivos es aportar a la atención de la seguridad alimentaria y nutricional en etapas claves de la vida [12].
La TUS es una tarjeta prepaga, con formato de banda magnética, utilizada en más de 600 comercios solidarios para la adquisición de alimentos, artículos de higiene personal y del hogar, así como productos y servicios que contribuyen al proceso de inclusión e integración social, permitiendo la autonomía del beneficiario en la selección de los productos de acuerdo a sus necesidades. La población objetivo del programa son los hogares en estado de indigencia o vulnerables al mismo, priorizando aquellos hogares con menores de 18 años.

La TUS es intransferible. Sólo puede ser utilizada por el titular, quien debe presentar su cédula de identidad al realizar las compras. Existen casos excepcionales en los que el MIDES extiende una carta poder para que otra persona pueda utilizar la tarjeta por un período de seis meses. El saldo cargado mes a mes en la TUS depende del número de menores de 18 años y embarazadas que vivan en su hogar [14].
A continuación se definen conceptos del programa que son utilizados a lo largo del documento.

TUS
Tarjeta Uruguay Social.

Comercio Solidario
Comercios formalizados, de pequeño porte, habilitados para realizar transacciones con la TUS.

AFAM
Programa implementado por MIDES y BPS. Se trata de una prestación económica a familias en situación de vulnerabilidad socioeconómica [12].

SIIAS
Sistema Integrado de Información del Área Social. Proyecto gubernamental destinado a generar un registro único de personas beneficiarias de programas sociales.

BPS
Banco de Previsión Social.

4.2.1 Requerimientos primarios de análisis de datos

A continuación se detallan los requerimientos relevados con los usuarios en una primera instancia. Se pretende realizar un análisis de los datos del programa teniendo en cuenta aspectos relacionados con la emisión de las tarjetas, los productos que son comprados con ella y el gasto producido.

TUS-1 Carga de tarjetas
Se desea saber cuántas tarjetas se cargan mensualmente y el monto cargado en ellas. Interesa conocer esta información en función de la distribución territorial de los beneficiarios.

TUS-2 Tarjetas duplicadas
Se desea saber la cantidad de tarjetas duplicadas: distintas personas que utilizan la misma tarjeta.

TUS-3 Tarjetas utilizadas por terceros
Se desea saber la cantidad de tarjetas utilizadas por un no titular. Esta funcionalidad se piensa para los casos en los que se autoriza el uso a terceros por medio de cartas poder.

TUS-4 Cantidad de productos
Se desea conocer la cantidad de productos que se compran con la tarjeta, y el importe, por fecha, categoría, zona geográfica y por comercio.

TUS-5 Variación de precios
Se desea conocer la variación de precios de productos por fecha, zona y comercio.

TUS-6 Distribución del gasto
Se desea conocer cómo se distribuye el gasto de las tarjetas discriminados por fecha, zona geográfica, por producto o rubro, y por las características de los beneficiarios. En una primera instancia interesan los atributos sexo, situación laboral y nivel educativo, pero se desea que, en el futuro, se puedan agregar nuevos atributos para el análisis.

TUS-7 Gasto por fecha

Se desea saber en qué fecha del mes se registran los mayores gastos. También interesa conocer cómo se distribuye el gasto en los diferentes meses del año.

TUS-8 Diferencia entre carga y gasto
Interesa conocer la variación entre el dinero que se carga en las tarjetas y el que se gasta mensualmente.

4.2.2 Fuentes de datos

Los datos utilizados para alimentar el DW provienen de múltiples fuentes. La principal fue proporcionada por *Scanntech*, empresa proveedora del software a los comercios solidarios. Los datos corresponden a las compras realizadas por los beneficiarios de la TUS en dichos comercios.

La Figura 4-12 presenta un subconjunto del modelo relacional de la base de datos (*Oracle*) del sistema utilizado por *Scanntech*. El mismo contiene las estructuras necesarias para registrar todo lo concerniente a las compras de productos con la tarjeta. El modelo completo se muestra en el Anexo 3.

Cada fila de la tabla *MOVIMIENTOS* contiene los datos de un movimiento. Cada uno de ellos se identifica por un número de movimiento y el número de la empresa en el cual se registra. En el conjunto de datos proporcionado todos los movimientos corresponden a compras cuyo importe fue acreditado total o parcialmente utilizando la TUS.

Un movimiento contiene un conjunto de líneas de pago, cada una identificada por el número de movimiento, el número de la empresa y el número de línea. Por ejemplo: si parte de una compra se abona en efectivo, otra parte con tarjeta de crédito y otra parte con la TUS, entonces ese movimiento tendrá tres líneas de pago. La tabla que registra las líneas de pago se llama *MOVIMIENTOS_DE_PAGOS*.

Cada movimiento contiene un conjunto de líneas de producto comprado. Cada línea de producto es identificada por un número, además del número de movimiento y el número de la empresa. La tabla que registra las líneas de producto se llama *MOVIMIENTOS_DETALLES*. Cada fila de esta tabla contiene el detalle de un producto comprado, como el código del producto, su código de barras, su descripción y rubro; además de la cantidad de producto comprado y el importe abonado.

Los movimientos que tienen un pago con la TUS, tienen una fila con dicho pago en una tabla llamada *MI_MOVIMIENTOS_PAGOS*. Cada tupla de esta tabla se identifica con el número de

movimiento y el número de empresa. Contiene el número de cédula del comprador, el número de tarjeta y el importe abonado con la tarjeta.

En la representación obviamos los catálogos de empresa, sucursal, departamento y localidad, también proporcionados por la empresa de gestión.

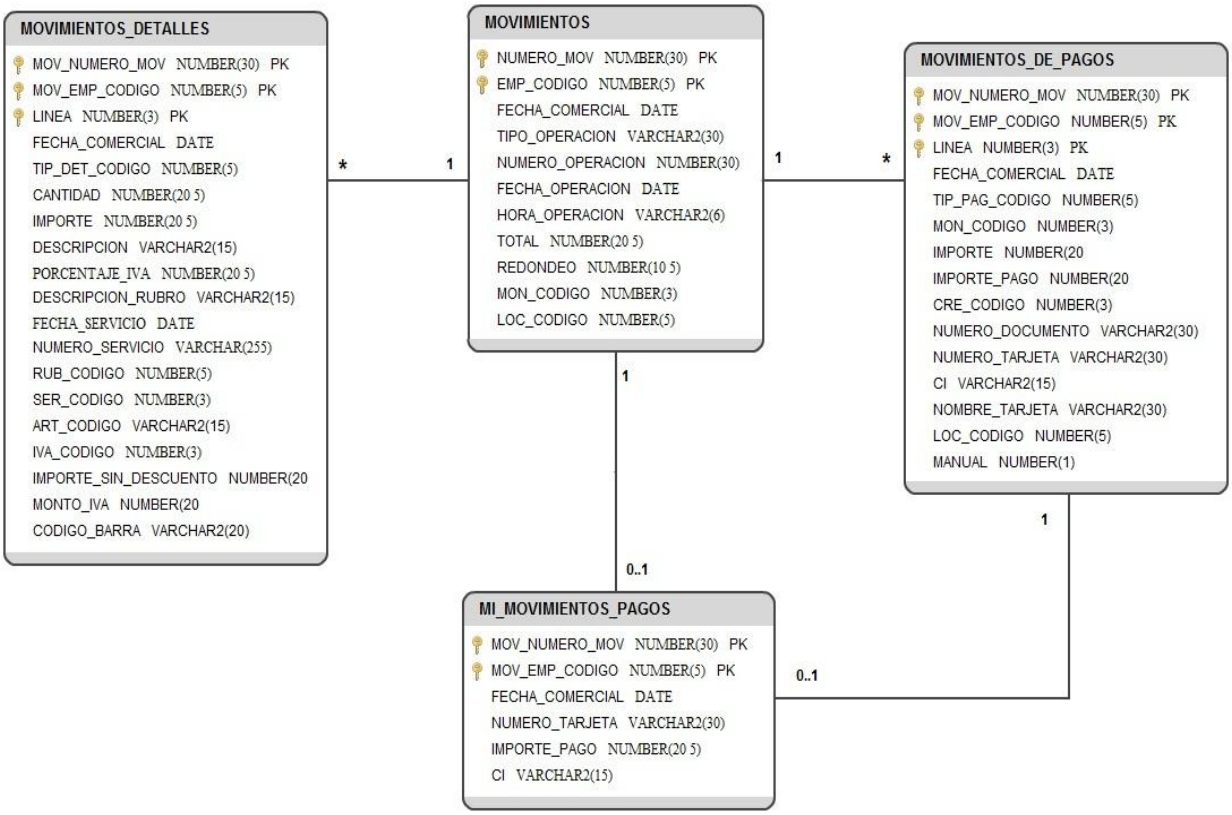


Figura 4-12: Modelo relacional de la base de datos de *Tarjeta Uruguay Social*

Con el fin de obtener datos complementarios para la implementación del DW, así como también para solucionar inconsistencias en los datos de la fuente principal, se utilizaron dos fuentes de datos auxiliares. Una de ellas consiste en archivos de texto con datos procesados de TUS, con los que MIDES alimenta a SIIAS, y de los cuales se extrajo la relación entre el número de tarjeta y el número de cédula del usuario. La otra fuente consiste en las tablas *InterfazPersonasNucleo* e *InterfazDocumentosPersonas*: subconjunto de la base de datos (*Sql Server*) de AFAM, proporcionadas por BPS e ilustradas en la Figura 4-13.

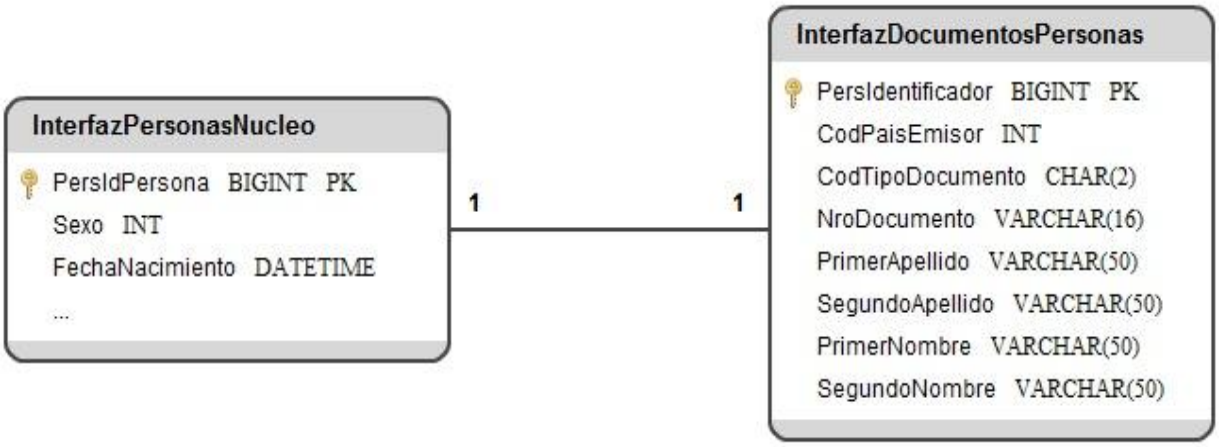


Figura 4-13: Tablas de Personas y Documentos de AFAM

4.2.3 Análisis de fuentes de datos y requerimientos

En esta sección se analiza cuánto se satisfacen los requerimientos en función de las fuentes de datos proporcionadas.

Al analizar la estructura de la fuente de datos principal, se observa que la misma no contiene la información necesaria para satisfacer los requerimientos TUS-1, TUS-2, TUS-3 y TUS-8. Con respecto a los requerimientos TUS-1 y TUS-8, no existe información acerca de la cantidad de tarjetas que fueron cargadas ni del monto cargado en cada una. Lo que se tiene es el gasto realizado para las tarjetas que fueron utilizadas.
Para el caso de los requerimientos TUS-2 y TUS-3, no se proporciona información de las cartas poder utilizadas para habilitar el uso de la TUS a una persona que no es la titular de la misma. Con las fuentes de datos provistas sólo se podría buscar compras de una misma tarjeta con distinta cédula. En conjunto con los usuarios se tomó la decisión de eliminar estos cuatro requerimientos de la lista.

Se observa que una misma compra puede contener muchas líneas de pago, sin que sea posible diferenciar los productos que fueron comprados con la TUS de los que fueron comprados utilizando otra forma de pago. Esto hace que sea imposible satisfacer el requerimiento TUS-4 tal como fue formulado originalmente. En cambio, sí se puede conocer los productos contenidos en una compra en la cual se utilizó la TUS, esto obliga a reformular el requerimiento.
Con respecto a la distribución del gasto se tomó la decisión de dividir en dos enfoques: uno que considera las compras realizadas en donde al menos una parte se pagó con la TUS, y el

otro que considera sólo el gasto hecho con la TUS (pagos parciales de las compras). De esta manera se satisface el requerimiento TUS-6 y el TUS-4 reformulado.
Los requerimientos TUS-5 y TUS-7 pueden ser satisfechos sin ser reformulados.

4.2.4 Requerimientos finales de análisis de datos

TUS-4 Cantidad de productos
Se desea conocer la cantidad de productos, y el importe de los mismos, de las compras realizadas total o parcialmente utilizando la TUS, por fecha, categoría, zona geográfica y comercio.

TUS-5 Variación de precios
Se desea conocer la variación de precios de productos por fecha, zona y comercio.

TUS-6 Distribución del gasto
Se desea conocer cómo se distribuye el gasto de las tarjetas discriminados por fecha, zona geográfica, producto o rubro, categoría de producto, y características de los beneficiarios. En una primera instancia interesan los atributos sexo, situación laboral y nivel educativo, pero se desea que, en el futuro, se puedan agregar nuevos atributos para el análisis.

4.2.5 Diseño conceptual

Aquí se presenta el modelo conceptual del programa *Tarjeta Uruguay Social*. Se indican qué datos son interesantes y cómo se relacionan entre sí.
En primer lugar se identifican dimensiones y medidas, luego se muestra cómo éstas se relacionan para satisfacer los requerimientos. Luego, se exhiben las relaciones dimensionales y el cuadro de roll-up.

4.2.5.1 Dimensiones

A continuación se describen las dimensiones del programa Tarjeta Uruguay Social.

- **Tiempo**
  La dimensión Tiempo se utiliza para el análisis del uso de la tarjeta desde una perspectiva temporal. Permite analizar las compras por fecha.
  Esta dimensión se estructura en una jerarquía de tres niveles (Figura 4-14), correspondientes a Día, Mes, Trimestre y Año.

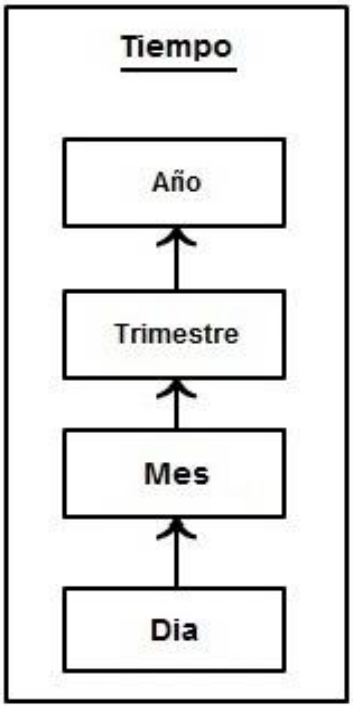


Figura 4-14: Dimensión Tiempo

- **Comercio**

  Esta dimensión se estructura en una jerarquía de tres niveles, en donde el nivel más granular es el local comercial en el cual se realiza el movimiento, que luego se agrupa en localidad y luego en departamento (Figura 4-15). Esta permite el análisis de las compras con la tarjeta desde un punto de vista geográfico.

Figura 4-15: Dimensión Comercio

- **Producto**

  La dimensión Producto identifica los productos comprados con la TUS. En los datos de origen estos productos pueden estar definidos o bien como un artículo o bien como un rubro. Un artículo representa a un producto específico mientras que el rubro es una generalización de los artículos.

  Ejemplo de artículos:

  - Arroz Saman 1 Kg.
  - Aceite Cocinero.
  - Manzana Roja.

  Ejemplos de rubro:

  - Frutas y Verduras.
  - Carnicería.

  Como se muestra en la Figura 4-16, los productos se agrupan en categorías, las cuales son definidas por el usuario según sus necesidades. Cabe destacar que en los datos de origen no existe ninguna categorización de los productos.

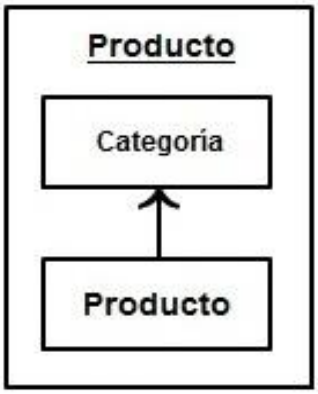


Figura 4-16: Dimensión Producto

- **Persona**

  La dimensión Persona (Figura 4-17) permite realizar un análisis en función de los datos del beneficiario de TUS. Algunos de estos datos son la fecha de nacimiento, el sexo, el nivel educativo y la situación laboral. Además, registra los números de tarjeta y la cédula de identidad de las personas.
  Esta dimensión está estructurada en varios niveles. El más granular corresponde a la persona, identificada por su cédula de identidad. El siguiente nivel se divide en varias jerarquías de agregación: una por cada atributo de análisis.

  En el sistema informático que se utiliza para el registro de las compras realizadas con la TUS, el numero de cédula es ingresado de forma manual por el cajero al momento del pago. La única información que se obtiene de forma automática es el número de la tarjeta, el cual la identifica. Por esto se decidió tener, en esta dimensión, el número de la tarjeta de cada persona como atributo. Esto ayuda a identificar en la dimensión, en el proceso de carga de la tabla de hechos, a la persona que realiza la compra.

  Se decidió definir a la persona como el nivel más fino de granularidad de la dimensión en lugar de agrupar los datos definidos en los requerimientos. Esta decisión se tomó por dos motivos: el primero es que los datos de las personas que fueron requeridos no están completos en las fuentes, por lo que es necesario actualizarlos periódicamente a medida que se vaya recolectando nueva información. Esta solución permite actualizar los datos de la dimensión sin tener que volver a cargar las tablas de hechos. Por ejemplo, si se lograse obtener una nueva fuente de datos con la información del sexo de los beneficiarios que no tienen ese dato, simplemente habría que actualizar la dimensión.

El otro motivo es prever la posibilidad de incorporar nuevos atributos, correspondientes a los beneficiarios, que resulten interesantes para el análisis demográfico y que no hayan sido contemplados en los requerimientos iniciales.
De esta forma se tiene flexibilidad ante la eventual modificación de requerimientos. Por ejemplo: si se quisiera medir las compras en función de la ascendencia de las personas, alcanzaría con agregar el atributo *Ascendencia* a la dimensión *Persona* sin tener que regenerar toda la tabla de hechos.

Problema con datos Cambiantes

Algunos datos acerca de las personas pueden cambiar con el tiempo. Esto nos enfrenta al problema de dimensión cambiante, o *Slowly Changing Dimension (SCD)* [6]
En esta dimensión podríamos clasificar los atributos que cambian en dos tipos:

Tipo Uno

Son los atributos de los cuales no interesa mantener un histórico. El valor que se tiene en cuenta es el actual y no el valor al momento en que se consumó el hecho. Un ejemplo de dato que pertenece a este tipo es el nombre de la persona, debido a que al momento de realizar el análisis interesa conocer el nombre actual. Otro ejemplo es la fecha de nacimiento, cuyo cambio sólo puede producirse a causa de un error al momento del ingreso.
Estos atributos se enmarcan dentro de la clasificación de una *SCD* de tipo I, descripta en [6]. Allí se recomienda sobrescribir el valor antiguo del atributo con el valor nuevo, sin mantener ningún histórico al momento de actualizar la dimensión.

Tipo Dos

Son los atributos de los cuales interesa conocer el valor del dato al momento en que se consumó el hecho y no su valor actual, por lo que los cambios en estos atributos no se pueden solucionar sobrescribiendo su valor, porque en ese caso se perdería la información histórica y el análisis de los hechos anteriores al cambio quedaría erróneo. Dentro de este tipo se encuentran los datos de la persona de los que, en los análisis estadísticos, interesa conocer su valor al momento de la compra y no al momento actual. Como ejemplo supongamos que una persona *A*, que tenía 10 hijos en el año 2010, tiene un nuevo hijo en el año 2012. ¿Cómo se contabilizarían, en el año 2012, los movimientos que dicha persona realizó durante el año 2010, como correspondiente una persona con 11 hijos o como una persona con 10 hijos? En este caso es deseable la segunda opción, que ese movimiento se contabilice como realizado por una persona con 10 hijos.
Si esta clase de datos se manejasen simplemente modificando el valor en la dimensión *Persona*, el análisis contaría ese movimiento teniendo en cuenta el valor

actual del dato. Se observa una dimensión cambiante. Para solucionar el problema se analizan las siguientes soluciones.

En primer lugar, considerando la recomendación descripta en [6], la dimensión *Persona* se podría manejar como una *SCD* de tipo II, la cual requiere agregar un nuevo registro cuando los datos de la persona cambien. A esta solución se le encontraron algunas dificultades en nuestro modelo. El primer problema tiene que ver con la dificultad para manejar el historial de una persona. Es preferible tener un sólo registro por persona en la dimensión para poder generar el historial de una forma sencilla.
Otro problema es el tamaño de la dimensión. Agregar un nuevo registro por cada cambio podría hacer que la misma creciera mucho, ya que se trata de una dimensión que contiene, en una primera versión, todos los datos de las personas beneficiarias de TUS (que representan unas 70.000 por mes), pero que debe ser escalable para agregar a las personas de otros programas. Por estas dificultades fue que se descartó tratar a sus atributos como de Tipo Dos.
Se decidió tomar la solución propuesta en [15]. Según las recomendaciones de esta publicación, se decide mantener en la dimensión *Persona* aquellos datos de los cuales no interesa mantener un histórico (los de tipo 1), y crear otra dimensión con los datos cambiantes de la persona, la cual es nombrada *Dimensión Demográfica* (Figura 4-18). Dicha dimensión contiene los datos correspondientes a la situación laboral y nivel educativo de las personas.
La dimensión *Persona* mantiene una referencia a la esta nueva dimensión que indica cuál es registro de que le corresponde al momento actual. De esta manera se mantiene la información actual de la persona. A su vez, cada registro en las tablas de hechos mantiene una referencia a la *Dimensión Demográfica* que indica cual era la información de la persona correspondiente al momento en el cual se consumó el hecho. Con esta solución se mantiene el histórico de las personas, se mantiene un sólo registro por persona y se obtiene la información demográfica al momento de la compra.
Cómo aspecto negativo se puede observar que esta solución no contempla el histórico de las distintas situaciones socioeconómicas por las que atraviesa una persona, y que se registran en la *Dimensión Demográfica*.

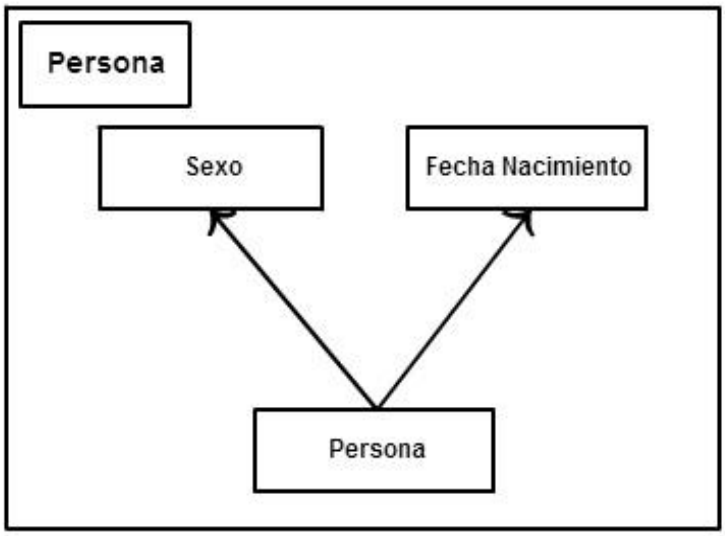


Figura 4-17: Dimensión Persona

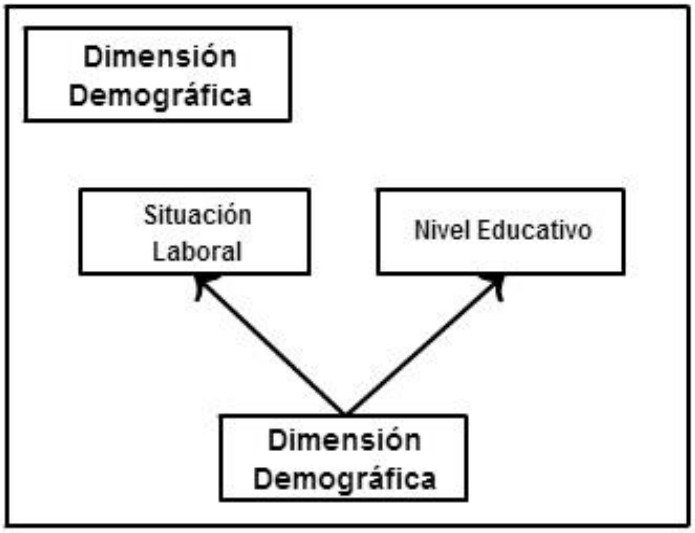


Figura 4-18: Dimensión Demográfica

- **Edad**

  Se decidió modelar la edad de una persona como una dimensión independiente debido a que se trata de un atributo que cambia todos los años y en el modelo interesa conservar la edad que tenía la persona al momento de la compra y no al momento de la consulta.

Se podría haber mantenido la edad como un atributo dentro de la dimensión demográfica, pero se decidió modelarla como una dimensión independiente a causa de la velocidad de cambio de este atributo. Esta velocidad podría hacer que la dimensión demográfica creciera rápidamente, y además se debería actualizar constantemente la dimensión *Persona* para modificar la referencia a la dimensión demográfica, y de esta manera mantener la información actual de la persona.
La dimensión *Edad* tiene dos niveles de agregación (Figura 4-19), en donde las edades se agrupan en diferentes rangos.

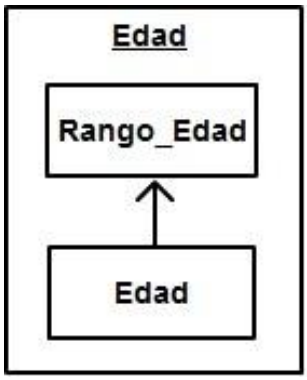


Figura 4-19: Dimensión Edad

4.2.5.2 Medidas

- **Cantidad**: Representa la cantidad de productos comprados.
- **Cantidad Movimientos**: Representa la cantidad de movimientos de las tarjetas.
- **Monto**: Representa la cantidad de los montos totales de las compras.
- **Monto Unitario**: Medida calculada que representa el cociente entre las medidas *Monto* y *Cantidad*.

4.2.5.3 Relación entre dimensiones, medidas y requerimientos

Esta sección muestra cómo se utilizan las diferentes dimensiones y medidas para satisfacer los requerimientos.

TUS-4 Cantidad de productos

- **Dimensiones:**
  - ***Tiempo:*** Interesa conocer la cantidad de productos, y el importe pagado por ellos, discriminados por fecha.
  - ***Producto:*** Interesa conocer la cantidad de productos, y el importe pagado por ellos, según su categoría.
  - ***Comercio:*** Interesa conocer la cantidad de productos, y el importe pagado por ellos, discriminados por comercio y zona geográfica de estos últimos.

- **Medidas**
  - ***Cantidad.***
  - ***Monto.***

TUS-5 Variación de precios

- **Dimensiones:**
  - ***Tiempo:*** Interesa conocer la variación de precios de productos según la fecha.
  - ***Producto:*** Interesa conocer variación de precios de productos.
  - ***Comercio:*** Interesa conocer variación de precios de productos según la zona y comercio.

- **Medidas**
  - ***Monto Unitario.***

TUS-6 Distribución del gasto

- **Dimensiones:**
  - ***Tiempo:*** Interesa conocer el gasto de las tarjetas según la fecha.
  - ***Producto:*** Interesa conocer el gasto de las tarjetas según los productos y sus categorías.
  - ***Comercio:*** Interesa conocer el gasto de las tarjetas según la zona y comercio.
  - ***Persona:*** Interesa conocer el gasto de las tarjetas según las características del beneficiario.

- **Medidas**

- ***Monto.***

TUS-7 Gasto por fecha

- **Dimensiones:**
  - ***Tiempo:*** Interesa conocer el gasto de las tarjetas según la fecha.

- **Medidas**
  - ***Monto.***

4.2.5.4 Relaciones dimensionales

Como se observó en la sección 4.2.5 (Análisis de fuentes de datos y requerimientos), con el modelo utilizado para guardar los datos de las compras con la TUS es imposible conocer, dado un movimiento, qué productos fueron pagados por la TUS y cuales con otros medios de pago (efectivo, crédito, etc). Lo que es posible conocer es en cuáles movimientos existe al menos un pago con la TUS.
Durante el análisis de los datos se observó que, en general, en los movimientos en los que existe al menos un pago con la TUS, el 95% del total del monto es pagado con la misma y el 5% con otro medio de pago. Teniendo en cuenta esta información, se decidió modelar dos relaciones dimensionales complementarias, una en la que se miden sólo los gastos realizados con la tarjeta y otra muy similar, pero en la que además se agrega la dimensión *Producto*. De esta manera, se pueden analizar los productos de compras en las cuales se utilizó la tarjeta.
En el modelo *Gasto* se consideran sólo los pagos con la TUS, así se conoce el dinero que realmente se gastó con cada tarjeta. Para esto se tuvo que suprimir la dimensión producto de este modelo.
El otro modelo, llamado *Gasto-Producto*, permite conocer qué productos fueron comprados en un movimiento. Para esto se cargan todos los productos comprados en los movimientos en los cuales al menos una parte se pago con la TUS, aunque el total del movimiento no haya sido acreditado con la tarjeta. Esto permite analizar qué productos compran los usuarios, y el error no sería mayor que el 5% según lo explicado anteriormente. Cabe destacar que la suma total del monto en los dos modelos será distinta. El segundo no refleja al 100% el gasto de la TUS, por lo que pueden aparecer productos prohibidos para la compra con la tarjeta.

A modo de ejemplo: supongamos que el Usuario U1 realiza una compra en la cual adquiere un producto P1 de valor $100 y un producto P2 de valor $150. Este cliente paga con la TUS la suma de $200 y contado los restantes $50.
En los datos origen esto quedaría representado como se muestra en la Figura 4-20.

Tabla MOVIMIENTOS

Comercio	Movimiento	Monto
C1	M1	250

Tabla MOVIMIENTOS_DE_PAGOS

Comercio	Movimiento	Usuario	Tipo Pago	Monto
C1	M1	U1	TUS	200
C1	M1	-	Contado	50

Tabla MOVIMIENTOS_DETALLES

Comercio	Movimiento	Producto	Monto
C1	M1	P1	100
C1	M1	P2	150

Figura 4-20: Ejemplar de datos de las fuentes de origen de TUS.

Al separar el análisis en estos dos modelos, si en el modelo *Producto* se analiza cuánto gastó el usuario U1, se sumará el total de los productos que este usuario compró, por lo que se resolvería que gastó $250, a diferencia de los $200 que se observaría con el modelo *Gasto-Producto*. La diferencia se encuentra en que el primer modelo permite analizar cuáles son los productos que el usuario compró.

Las Figuras 4-21 y 4-22 presentan ambos modelos descriptos.

- **Gasto**

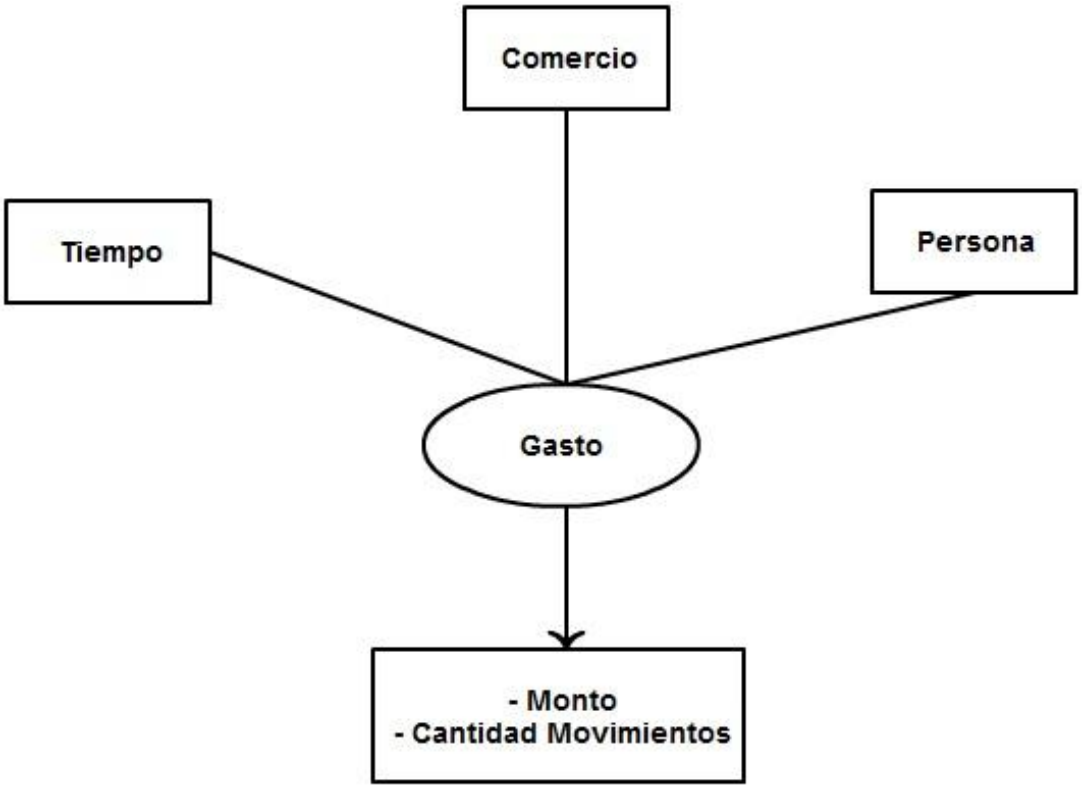


Figura 4-21: Relación Dimensional Gasto

- **Gasto - Producto**

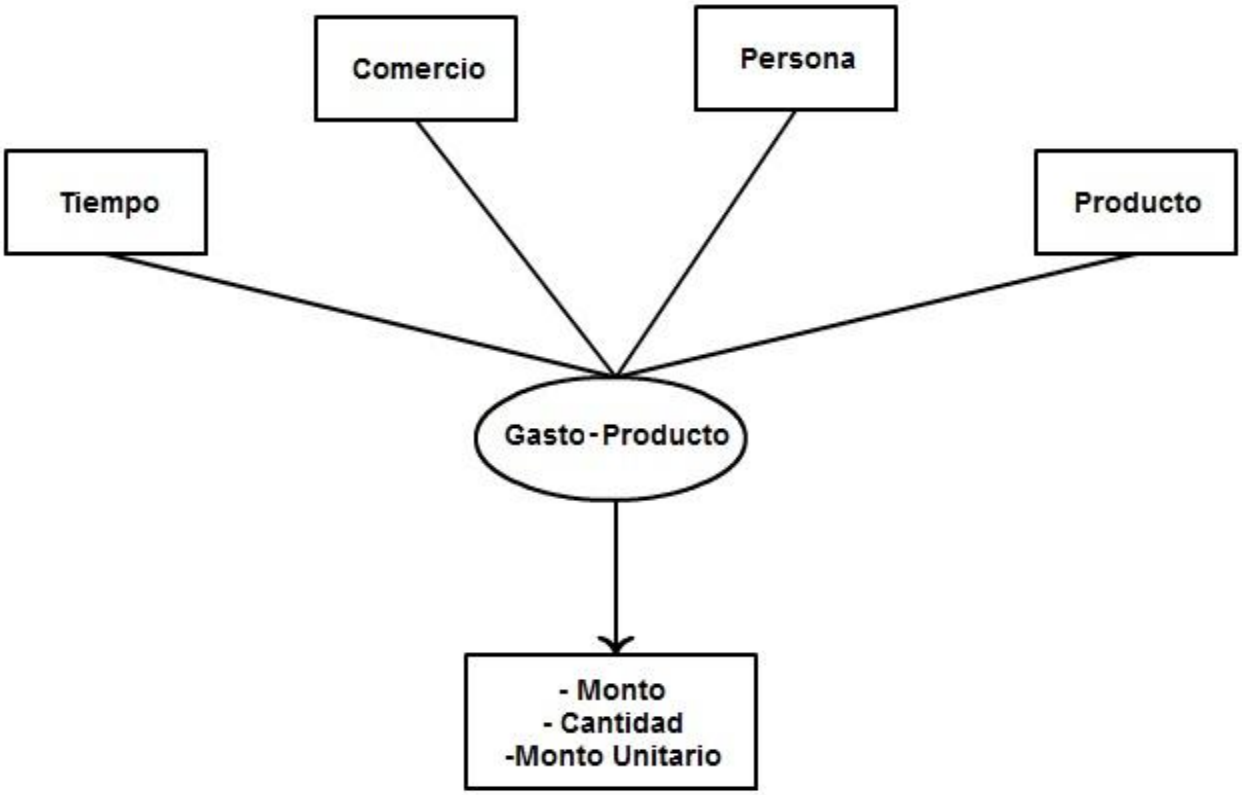


Figura 4-22: Relación Dimensional Gasto-Producto

4.2.5.5 Cuadro de Roll Up

A continuación se muestran los cuadros de Roll Up para los modelos dimensionales del programa TUS.

- **Producto**

		Cantidad	Monto	Monto Unitario
Tiempo	**Dia→Mes**	**+**	**+**	1
	Dia→Mes→Año	**+**	**+**	1
Comercio	**Comercio→Localidad**	**+**	**+**	1
	Comercio→Localidad→Departamento	**+**	**+**	1
Producto	**Producto→Categoría**	**+**	**+**	1
Persona	**Persona →Sexo**	**+**	**+**	1
Persona	**Persona →Fecha Nacimiento**	**+**	**+**	1
Edad	**Edad→RangoEdad**	**+**	**+**	1

[1] Se calcula después de agregar. Primero se agregan las medidas monto y cantidad, y luego se realiza el cociente entre ellas.

- **Gasto**

		Cantidad Movimientos	Monto
Tiempo	**Dia→Mes**	**+**	**+**
	Dia→Mes→Año	**+**	**+**
Comercio	**Comercio→Localidad**	**+**	**+**
	Comercio→Localidad→Departamento	**+**	**+**
Persona	**Persona→Sexo**	**+**	**+**
Persona	**Persona→Fecha Nacimiento**	**+**	**+**
Edad	**Edad→RangoEdad**	**+**	**+**

4.2.6 Diseño lógico

Para el diseño lógico de la solución se decidió, al igual que en el caso de *Uruguay Integra*, utilizar un *Star cluster schema*. La justificación de dicha decisión es la misma que para el caso anterior. Solamente se normalizó la jerarquía correspondiente a la ubicación geográfica (Localidad->Departamento), compartida por las dimensiones *DimComercio* y *DimDemográfica*.

La Figura 4-23 muestra el diseño lógico de la solución. Se crearon las tablas de hechos *FactProductos y FactGastos*, y las de dimensiones *DimComercio*, *DimPersona*, *DimEdad, DimDemográfica, DimLocalidad, DimTiempo* y *DimProducto*.

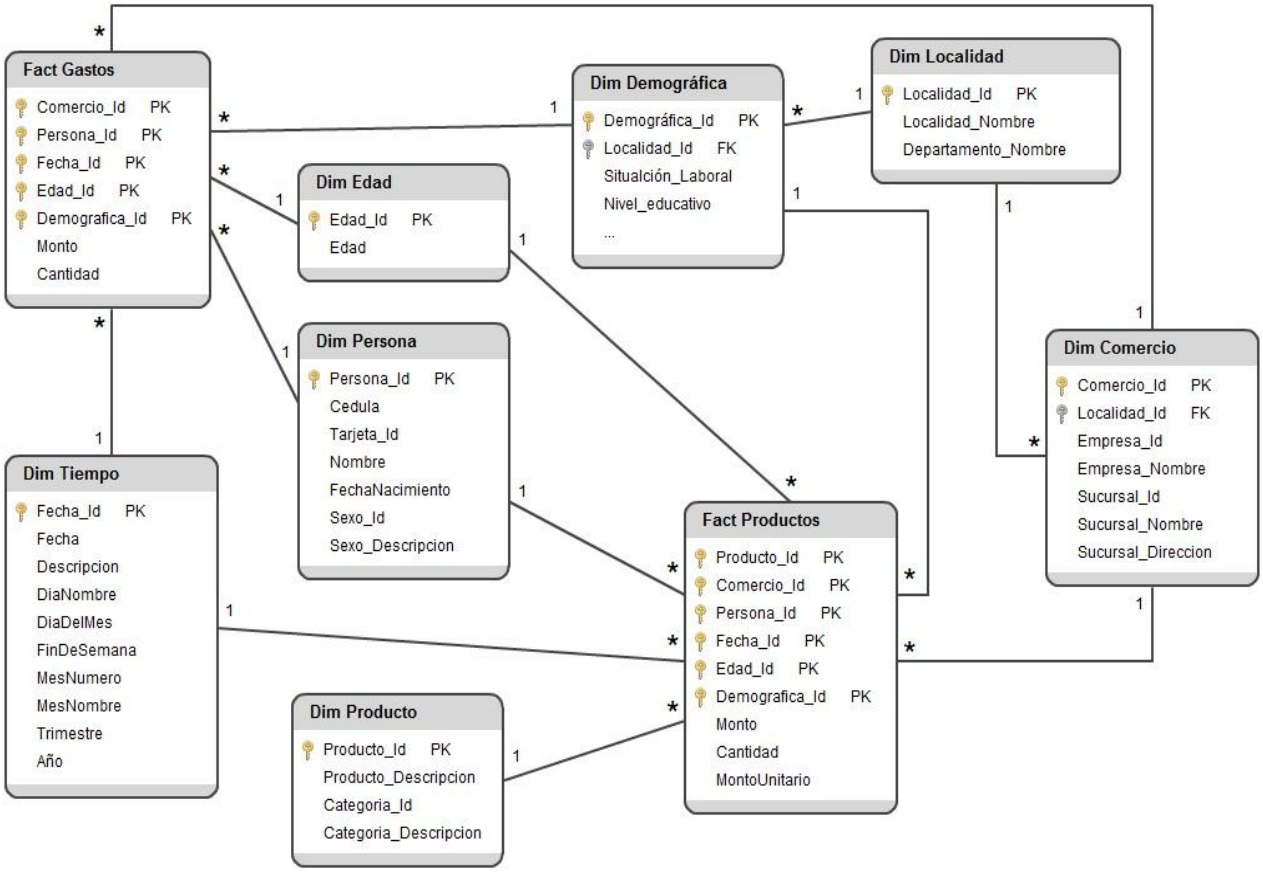


Figura 4-23: Diseño lógico de *Tarjeta Uruguay Social*

5 Procesos ETL

Este capítulo describe todo lo referente a la extracción, transformación y carga de datos desde las fuentes al DW.

5.1 Introducción

Los procesos de Extracción Transformación y carga (ETL) son los encargados de mantener los datos del DW actualizados.
La Figura 5.1 describe la arquitectura ETL utilizada en este proyecto. Se utiliza un ODS, desde donde se toman los datos y se procesan para almacenarlos en el DW. El procesamiento involucra la realización de tareas de integración, filtrado y limpieza de los datos. También se aplican transformaciones para agrupar y calcular nuevos datos.
En este trabajo se utilizaron dos herramientas ETL: *Integration Services* de la suite de *Microsoft* y *Kettle Data Integration* de *Pentaho*.

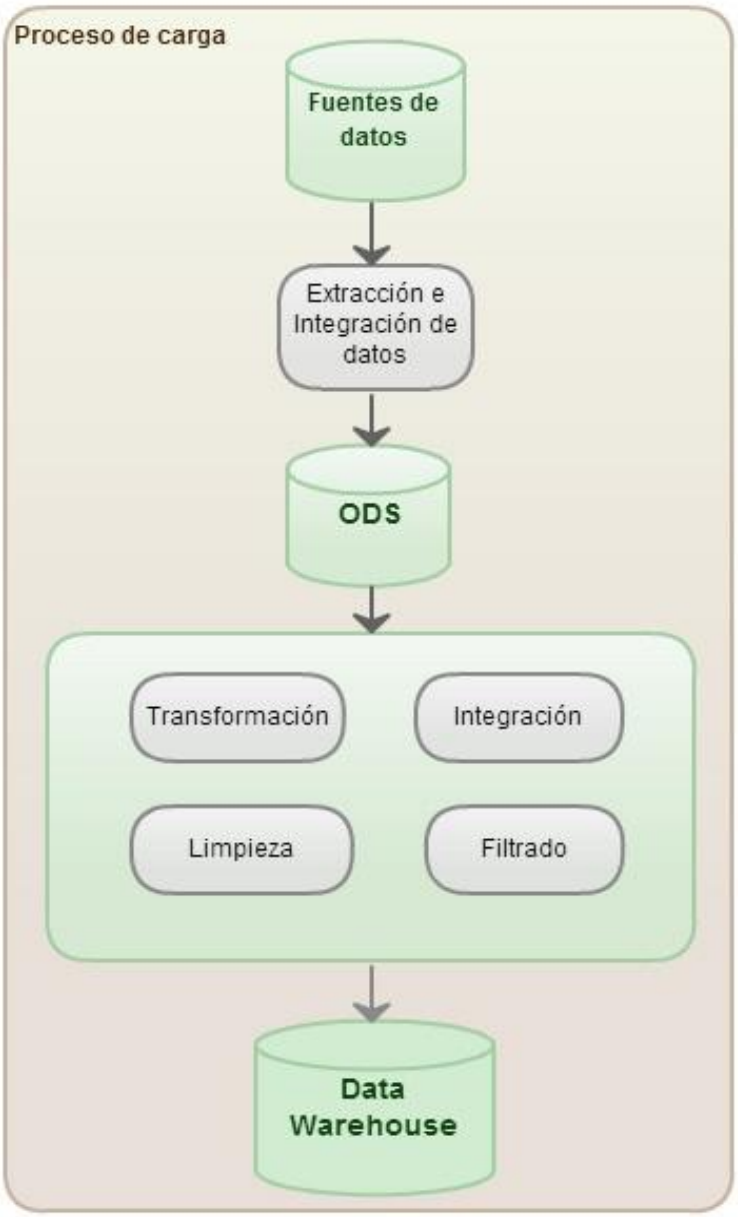


Figura 5-1: Arquitectura ETL

5.2 Limpieza de datos

El programa *Uruguay Integra* no cuenta con un Sistema de Información para el procesamiento de sus datos. Éstos últimos recorren un camino que comienza cuando son ingresados al sistema informático mediante herramientas de Data Entry, como *CSPro*, o formularios generados con el software de gestión *InnovaPortal*. En ninguno de los dos

casos se implementan las validaciones y restricciones necesarias sobre los formularios de ingreso, lo que origina que los datos se carguen con una gran cantidad de errores. Una vez que son ingresados, los datos son exportados a planillas Excel, siendo éstas últimas las fuentes de datos para la construcción del DW.
El programa *Tarjeta Uruguay Social* cuenta con un Sistema de Información que no contiene todos los datos necesarios para satisfacer los requerimientos del DW. Por tal motivo deben utilizarse fuentes de datos externas al sistema de gestión del programa.
Para este sistema no se cuenta con información acerca de las validaciones implementadas. Sin embargo, al realizar una inspección de los datos, se puede observar que en los mismos no existen controles sobre las cédulas que se ingresan.

5.2.1 Calidad de datos

Para el análisis de la calidad de las fuentes de datos se utilizó el enfoque *Data Profiling*, que consiste en el análisis de las instancias individuales de los atributos con el fin de obtener propiedades de los mismos. Se analiza su tipo de datos, longitud, contenido, rango de valores de dominio, unicidad y ocurrencia de valores nulos [9].
En nuestro caso, un procedimiento utilizado para la detección de errores sobre las planillas Excel de *Uruguay Integra* fue el ordenamiento de los valores por cantidad de ocurrencias, de manera de que los datos erróneos aparecieran contiguos a los correctos y, de esta forma, facilitar la comparación visual. Así también se identificaron valores faltantes y duplicados.
En el caso de *Tarjeta Uruguay Social* se realizó una inspección sobre las fuentes, que consistió mayormente en consultas de base de datos.
En ambos casos se encontraron problemas relacionados con las dimensiones *Completitud,* en su factor *Densidad*, y *Exactitud*, esta última en su factor *Correctitud Sintáctica*.

5.2.1.1 Dimensión de calidad: Completitud

Uruguay Integra
Se observó la presencia de valores nulos en los campos *fecha de nacimiento*, *localidad*, *cédula* (de los técnicos) y *nivel educativo*. En estos casos se optó por tener un código especial único (número 99) para indicar la ausencia de dato.

Tarjeta Uruguay Social
Se observó que la descripción de los artículos vendidos, campo utilizado en el proceso de identificación de los mismos, se encuentra truncada en quince caracteres. Ante la dificultad para obtener la descripción completa de los artículos, se decidió trabajar utilizando la descripción truncada.

En el requerimiento **TUS-6** se desea conocer la distribución del gasto de las tarjetas según el sexo, la situación laboral y el nivel educativo de los beneficiarios, además de contemplar la posibilidad de agregar nuevos atributos cuando estos estén disponibles. La Figura 5.2 muestra el porcentaje de personas a las que se les completaron los distintos atributos:

Atributo	*Porcentaje*
Sexo	90%
Fecha Nacimiento	90%
Nivel Educativo	85%
Situación Laboral	87%

Figura 5-2: Porcentaje de personas a las cuales se les completaron los atributos contemplados en el requerimiento TUS-6.

Al igual que en el caso de UI, se definió el valor *Sin dato* para indicar la ausencia de dato.

5.2.1.2 Dimensión de calidad: Exactitud

Uruguay Integra

Se observó en los datos la presencia de una gran cantidad de errores sintácticos, producidos ya sea por errores de digitación durante su ingreso o simplemente a causa de un libre albedrío a la hora de elegir valores de un conjunto dominio, esto último debido a la falta de controles en los formularios de ingreso de datos. Se encontraron nombres de una misma instancia, en entidades elegidas como dimensión, escritos de muchas formas diferentes. Las entidades que presentaron este problema fueron: *OSC, Grupo, Localidad, Motivos de abandono, Nivel educativo y Situación laboral*. Para solucionar este problema se implementaron diccionarios representantes del conjunto de valores posibles para las entidades, y se chequeó la existencia de dichos valores utilizando un algoritmo proporcionado por la herramienta ETL utilizada. Tanto el algoritmo como los diccionarios serán explicados en las secciones 5.2.2 y 5.2.3 respectivamente.

En el archivo fuente correspondiente a los datos de los técnicos del programa se encontraron números de cédula incorrectos. Éstos fueron detectados mediante la utilización del algoritmo de validación que compara el número de cédula con su dígito verificador (ver sección 5.2.2). Los números incorrectos fueron chequeados uno por uno en forma particular consultando a los usuarios.

Se hallaron errores de estandarización, como por ejemplo los campos *sexo* y *fecha* con diferentes formatos. Para solucionar el problema se decidió convertir los valores de las instancias de las entidades que presentaban este problema a un formato común:

- Todas las fechas se convirtieron al formato dd/mm/aaaa.

- Los valores de tipo cadena de caracteres alfabética fueron convertidos a mayúscula.
- A los números de cédula se les quitó puntos y guiones, y se los convirtió a tipo entero, incluyendo el dígito verificador.
- Los valores del campo *sexo* fueron unificados a la siguiente codificación:
    - 1 Masculino
    - 2 Femenino
    - 99 Sin Dato

Tarjeta Uruguay Social

El mayor problema encontrado fue la ausencia de un catálogo o lista de productos vendidos en los comercios solidarios. La solución a este problema representó una de las mayores dificultades del desarrollo del DW. Dicha solución consistió en la construcción de un diccionario de productos, el cual se explicará en forma detallada en la sección 5.3.3.3, que trata acerca de la carga de la dimensión *Producto.*

Se encontraron errores en los códigos de barras de los artículos, ya sea debido a ingresos manuales o codificaciones erróneas al ingresarlos en el sistema. Para detectar estos casos aplicamos el algoritmo de validación de códigos de barra (ver sección 5.2.2). Para los casos en que el algoritmo determina que el código de barras del artículo no es válido, se procede a la asignación de un código por defecto, que no es tomado en cuenta a la hora de identificar al producto.

Se constató que el formato del atributo correspondiente a las cédulas de identidad de los beneficiarios no se ajusta, en la mayoría de los casos, a la nomenclatura estándar de siete u ocho dígitos sin puntos ni guiones. Los datos presentan diferentes formatos, lo que dificulta la identificación del número de documento y su correspondiente validación. La solución a este problema se detalla en la sección 5.3.3.4, en donde se trata la carga de la dimensión *Persona.*

Modelo de calidad aplicado

En la Tabla 5.1 se presentan las dimensiones, con sus factores, y las acciones aplicadas a cada dato o conjunto de datos, para la evaluación y limpieza de los mismos

Programa	Campos	Completitud (Densidad)	Exactitud (Correctitud sintáctica)
UI	Fecha de nacimiento, Localidad y Nivel educativo	Para los valores nulos se adjudicó el código 99 que indica la ausencia de dato.	
	OSC, Grupo, Localidad, Motivos de abandono, Nivel educativo y Situación labora		Se construyeron diccionarios para cada una de las entidades.
	Cédula de técnicos		Se aplicó algoritmo de validación de cédulas. Las cédulas que no validaron fueron chequeadas en forma particular.
	Sexo y Fecha de nacimiento		Se realizó una estandarización con el fin de llevar los valores de los campos a un formato común.
TUS	Producto	Se trabajó con las descripciones incompletas.	Construcción de diccionario de productos.
	Sexo, Fecha de nacimiento, Nivel Educativo y Situación Laboral	Para los valores nulos se adjudicó el código 99 que indica la ausencia de dato.	
	Código de barras		Se aplicó el algoritmo de validación de códigos de barra. A los códigos que no validaron se le asignó un código por defecto.
	Cédula		Se unificó el formato. Se mantuvo el número de cédula sin dígito verificador.

Tabla 5.1: Modelo de calidad aplicado.

5.2.2 Algoritmos

En esta sección se explican los algoritmos utilizados para la limpieza de datos.

Compute Distance

Los algoritmos de *Compute Distance* están basados en la *Distancia de Levenshtein*, y se utilizan para determinar qué tan similares son dos cadenas de caracteres.

Se llama *Distancia de Levenshtein* al número mínimo de operaciones que son requeridas para transformar una cadena de caracteres en otra, en donde una operación consiste en una inserción, eliminación o sustitución de un carácter. Por ejemplo: la *Distancia de Levenshtein* entre las cadenas "MARIO" y "MARTA" es 2, porque se necesitan dos operaciones para transformar una en otra: la sustitución de la letra "I" por la letra "T", y la sustitución de la letra "O" por la letra "A"; mientras que la distancia entre "MARIO" y "MARTHA" es 3, porque a las sustituciones anteriores se le suma la inserción de la letra "H".

En este proyecto se utilizó la implementación concreta brindada por la transformación *Fuzzy LookUp,* de la herramienta ETL *Integration Services*. La misma compara una cadena de caracteres de entrada con cadenas de una tabla de referencia. Utiliza un algoritmo de *Compute Distance* que devuelve un resultado entre 0 y 1 que indica el grado de similitud de dos cadenas comparadas.

Validación de cédulas

Para corroborar que un número de cédula de identidad sea correcto se implementó el algoritmo definido por la Dirección Nacional de Identificación Civil (DNIC), el cual se basa en la comparación de los primeros dígitos del número de la cédula con su dígito verificador. Para mostrar el algoritmo tomaremos como ejemplo la cédula genérica $d_1d_2d_3d_4d_5d_6d_7-d_8$.

El algoritmo calcula el dígito verificador mediante la siguiente fórmula:

$$dv = \Big(10 - \big((d_1 * 2 + d_2 * 9 + d_3 * 8 + d_4 * 7 + d_5 * 6 + d_6 * 3 + d_7 * 4) \bmod 10\big)\Big) \bmod 10$$

Si el dígito calculado (dv) coincide con el dígito verificador (d_8), entonces la cédula es correcta.

Validación de códigos de barras

El código de barras consiste en un sistema basado en la representación de líneas verticales paralelas, de distinto grosor, que codifican una determinada información. Este sistema se utiliza para la identificación de un artículo, el cual se puede identificar de forma global o dentro de una cadena de comercios. Los códigos que se utilizan dentro de una cadena de comercios comienzan con el digito 2. Dentro de los diferentes grupos de códigos de barras

existentes se encuentra el *European Article Number* (*EAN*), adoptado por más de cien países y más de un millón de empresas en todo el mundo. El código utilizado para la identificación de los artículos en los comercios solidarios es el *EAN13*, que consta de trece dígitos estructurados de la siguiente forma: los tres primeros corresponden al código del país en donde reside la empresa, los siguientes cuatro o cinco dígitos corresponden al código de la empresa. La codificación del producto completa el código hasta el duodécimo dígito. El restante, llamado dígito de control, se utiliza para verificar la correctitud del código.
Para validar un código de barras *EAN13* se calcula el dígito de control a partir de los primeros doce dígitos del código. Para mostrar el algoritmo tomaremos como ejemplo el código genérico $d_1d_2d_3d_4d_5d_6d_7d_8d_9d_{10}d_{11}d_{12}d_{13}$. El algoritmo calcula el dígito de control mediante la siguiente fórmula:

$$dc = \left(10 - \left(\left((d_{12} + d_{10} + d_8 + d_6 + d_4 + d_2) * 3 + (d_{11} + d_9 + d_7 + d_5 + d_3 + d_1)\right) mod\ 10\right)\right) mod\ 10$$

Si el dígito calculado (dc) coincide con el dígito de control (d_{13}), entonces el código de barras es válido.

5.2.3 Diccionarios

Para enfrentar los problemas de Exactitud, en su factor Correctitud Sintáctica, se utilizaron diccionarios contenedores de los conjuntos dominio de las entidades del mundo real. Cada vez que se realizan cargas o actualizaciones al DW, los valores de las entidades que pueden contener errores se chequean contra los diccionarios. Éstos se van extendiendo a medida que aparecen nuevos valores que extienden los dominios.
Hay casos en los que existen diferentes valores del dominio que representan a la misma instancia de una determinada entidad, por ejemplo cuando un nombre es representado con una sigla o acrónimo. Como ejemplo concreto se puede mencionar que *Asociación de mujeres rurales uruguayas* y *A.M.R.U.* representan la misma instancia de la entidad OSC en el programa *Uruguay Integra*. En estos casos se optó porque todas las descripciones posibles de una instancia estuvieran contenidas en el diccionario que les corresponde, con la particularidad de que cuando se realiza una búsqueda por alguna de ellas, se obtiene como resultado el mismo identificador. En los casos en que la diferencia obedece simplemente a errores sintácticos, se utilizan algoritmos de *Compute Distance* (explicados en la sección anterior) para determinar si el valor ya pertenece al diccionario o debe ser ingresado como una nueva instancia. Las entidades para las que fue necesario crear un diccionario fueron: *Grupo, OSC, Motivos de abandono, Nivel educativo* y *Situación laboral* para el programa *Uruguay Integra*; y *Producto* y *Rubro* para el programa *Tarjeta Uruguay Social*.

La Tabla 5-2 muestra un ejemplo del diccionario correspondiente a la entidad *OSC*. El mismo consiste en una tabla con tres columnas: la primera contiene la clave de la instancia en el diccionario, la segunda contiene la descripción de la instancia y la tercera su identificador en la dimensión *OSC*. En este caso se muestra el ejemplo mencionado antes. Cuando se consulte al diccionario, tanto para *Asociación de mujeres rurales uruguayas* como para *A.M.R.U.,* se devolverá el identificador 15, esto indica que representan la misma instancia de *OSC*. Si la consulta al diccionario es por la descripción *Casa de la mujer de la Unoin*, que contiene un evidente error en su última palabra, el algoritmo de *Compute Distance* detectará que se trata en realidad de la instancia *Casa de la mujer de la Unión*, y se devolverá el identificador 16.

Clave	Descripción	Identificador
1	Asociación de mujeres rurales uruguayas	15
2	A.M.R.U	15
3	Casa de la mujer de la Unión	16

Tabla 5-2: Ejemplar del diccionario de la entidad OSC

5.3 Carga y actualización

Esta sección trata acerca de la carga y actualización de las tablas de hechos y las dimensiones del DW para los dos programas sociales implementados. En primer término se presentan flujos genéricos utilizados para la carga de las tablas de hechos y para las dimensiones que necesitan un diccionario. Luego se detallan los casos particulares de la carga de las tablas de hechos y las dimensiones de ambos programas.

La frecuencia de actualización varía para cada programa. *Uruguay Integra* se actualiza anualmente, mientras que *Tarjeta Uruguay Social* lo hace en forma mensual.

5.3.1 Flujos de carga genéricos

Tablas de hechos

La Figura 5-3 describe el flujo de carga genérico utilizado para realizar la carga y actualización de las tablas de hechos. Primero se extraen los datos de las fuentes y son normalizados aplicando operaciones sobre los datos que son cadenas de caracteres, como pasarlos a mayúscula y sacarle los espacios en blanco al principio y al final de la cadena.

Luego, para cada tupla de datos, se buscan en los diccionarios los valores correspondientes a las dimensiones con el fin de obtener sus identificadores. Luego se valida que las dimensiones obtenidas del diccionario estén cargadas en las tablas de dimensiones y se calculan las medidas. Finalmente, se ingresan los valores a la tabla de hechos. Si algún valor correspondiente a las dimensiones no es encontrado en el respectivo diccionario, o algún valor de las dimensiones, referenciadas en los datos a cargar en la tabla de hechos, no están en la tabla de dimensiones, la tupla se vuelca hacia el archivo log de inconsistencias para su posterior chequeo manual.

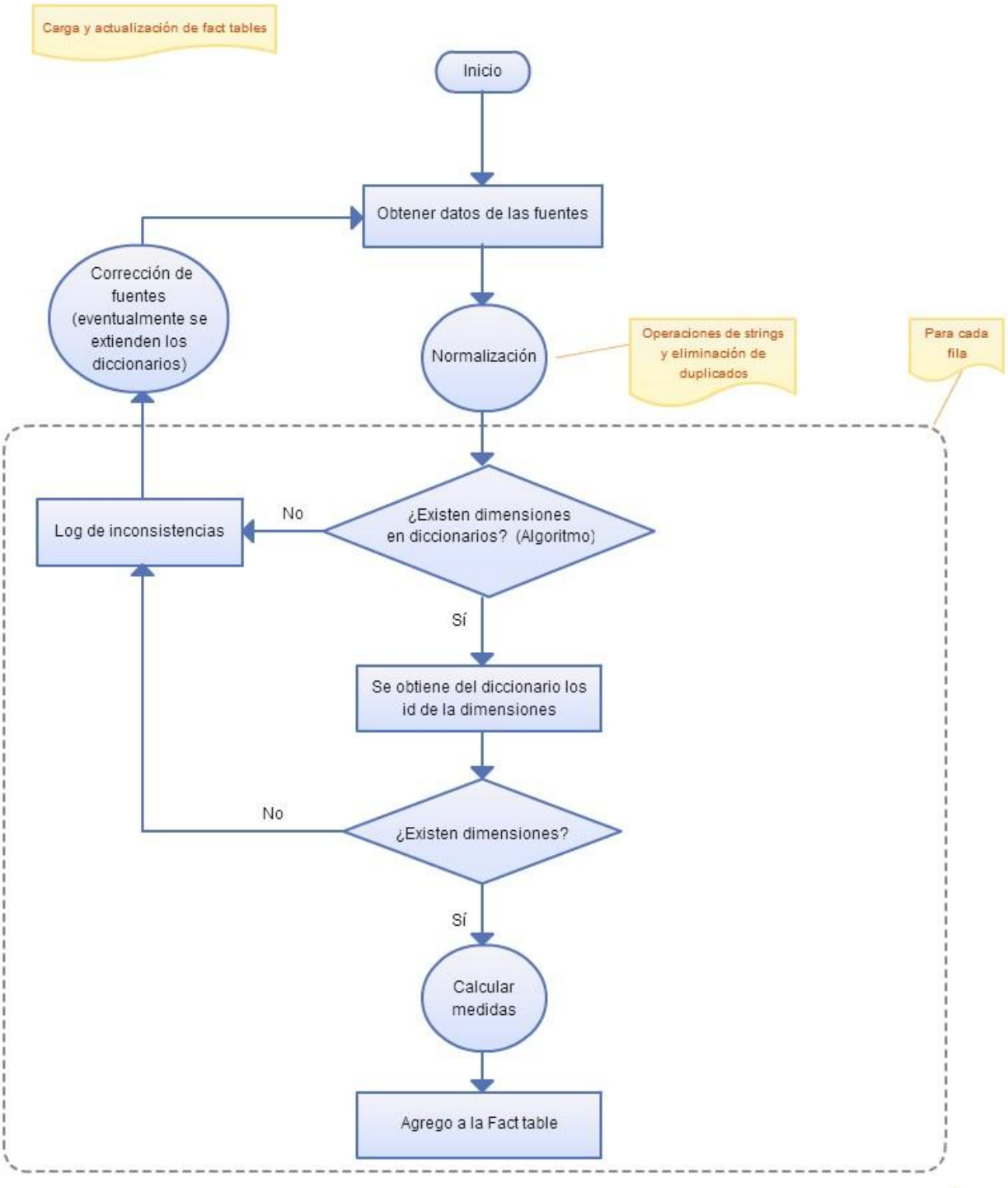


Figura 5-3: Flujo de carga y actualización de tablas de hechos

Dimensiones con diccionario

La Figura 5.4 describe el flujo utilizado para la carga inicial y actualización de las dimensiones cuyas entidades necesitaron la construcción de un diccionario.

Al igual que en la carga de las tablas de hechos, se comienza con la normalización de los datos. Luego, para cada valor, se realiza una búsqueda de comparación en el diccionario correspondiente. Si el valor es encontrado en el diccionario, y no pertenece ya a la dimensión que se está cargando, se ingresa como nuevo valor en la tabla de dimensión. Si, por el contrario, el valor no es encontrado en el diccionario, se vuelca hacia un archivo log de inconsistencias para que luego sea chequeado manualmente.

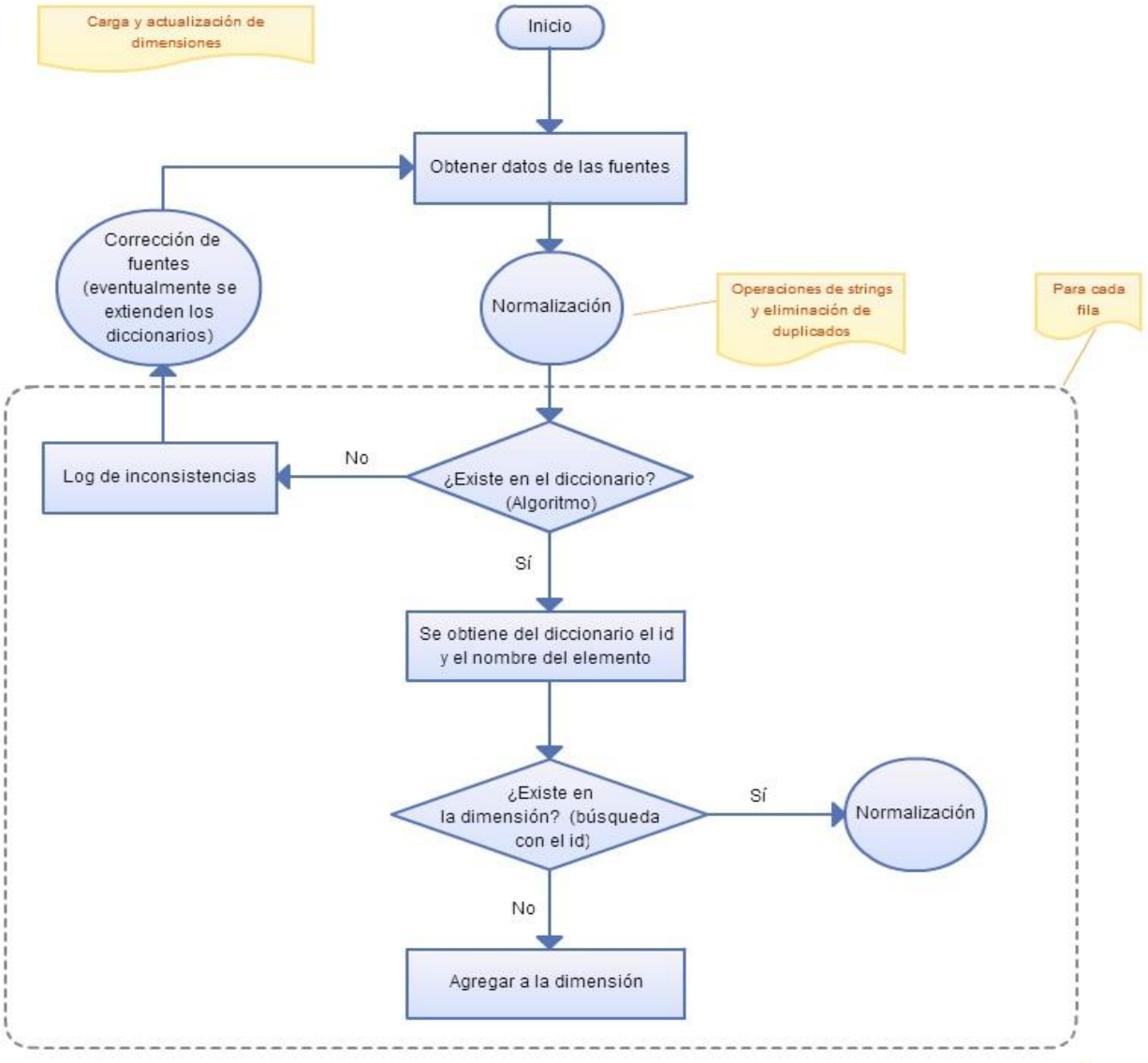


Figura 5-4: Flujo de carga y actualización de dimensiones

5.3.2 Carga de Uruguay Integra

En esta sección se muestran los flujos de carga para las dimensiones y tablas de hechos del programa *Uruguay Integra*.

5.3.2.1 Dimensiones

A continuación se muestra el flujo de carga para la dimensión *OSC*. Los correspondientes a las dimensiones *Localidad*, *Motivo* y *Grupo* se muestran en el Anexo 2. Las dimensiones Edad, Nivel Educativo y Situación Laboral se cargan en forma manual.

OSC

La Figura 5.5 muestra el flujo de carga para la dimensión *OSC*.

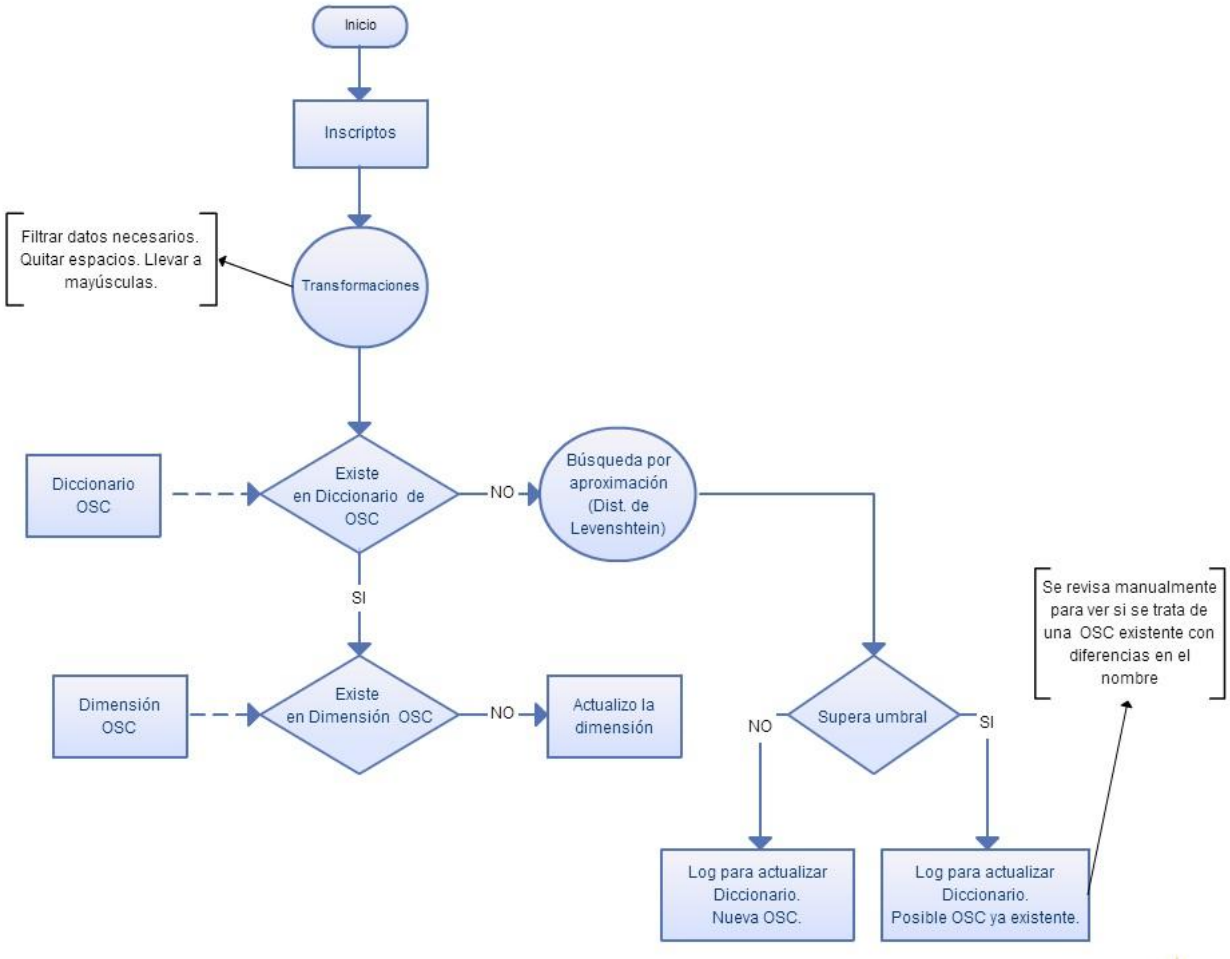


Figura 5-5: Flujo de carga: dimensión *OSC*.

5.3.2.2 Tablas de hechos

A continuación se muestra el flujo de carga para la tabla de hechos de *Contexto*. Los correspondientes a *Recursos* y *Producto* se muestran en el Anexo 2.

Contexto

La Figura 5.6 muestra el flujo de carga para la tabla de hechos *Fact Contexto*.

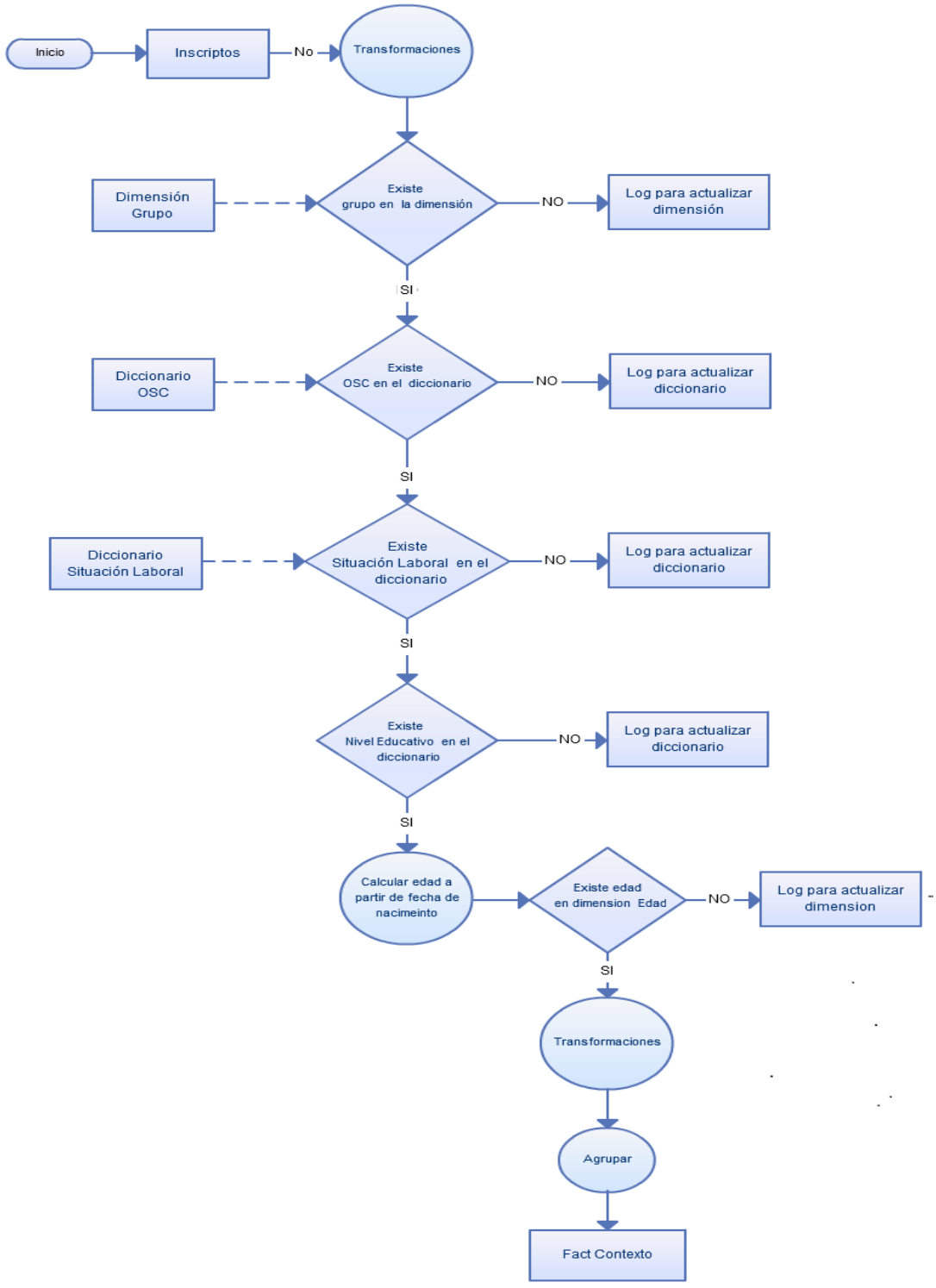


Figura 5-6: **Flujo de carga: Tabla de hechos *Fact Contexto.***

5.3.3 Carga de Tarjeta Uruguay Social

En esta sección se detallan las cargas de las dimensiones *Tiempo*, *Producto* y *Persona*, y las tablas de hechos de *Productos* y *Gastos*. El flujo de carga de la dimensión Comercio se muestra en el Anexo 2.

5.3.3.1 Dimensión Tiempo

Conforme a las necesidades del análisis, se decidió diseñar una tabla de dimensión *Tiempo* con granularidad a nivel de día; es decir, con un registro para cada día del año. La clave subrogada de la tabla consiste en un número entero con el formato yyyymmdd, que refiere a la concatenación de los números enteros que representan el año (yyyy), el mes (mm), y el día (dd) de la fecha correspondiente [16]. Además de la clave subrogada, cada fila de la tabla contiene los siguientes campos:

- Fecha en tipo *Datetime* de *SQL Server.*
- Descripción de la fecha en formato texto
- Descripción del día (Domingo, Lunes, etc).
- Número del día dentro del mes.
- Indicador para saber si el día corresponde a un fin de semana.
- Número de mes.
- Descripción del mes.
- Número correspondiente al trimestre.
- Año.

Este diseño presenta la ventaja de ser muy flexible debido a que en la tabla se pueden representar simultáneamente a todos los grupos de tiempo útiles para el análisis.
Los campos especiales de navegación, como, por ejemplo, el indicador de fin de semana, permiten definir lapsos arbitrarios de tiempo. Además, la posibilidad de incorporar fácilmente nuevos campos de navegación hace que la dimensión sea escalable.

La carga se realizó utilizando un script, en lenguaje SQL, en el cual se implementa un bucle cuyas iteraciones realizan la carga de un determinado rango de fechas.

5.3.3.2 Dimensión Producto

Según las características pretendidas del análisis de los datos del programa *Tarjeta Uruguay Social, Producto* resulta una de las dimensiones más importantes en el diseño del DW. Se pretende medir las cantidades de productos comprados y el dinero gastado en ellos. También se requiere que dichos productos estén categorizados. Sin embargo, las fuentes de datos con las que se cuenta no sólo carecen de la categorización requerida, sino que ni siquiera contienen un catálogo con la totalidad de los productos existentes en los comercios. La información acerca de los productos sólo puede obtenerse de la tabla *MOVIMIENTOS_DETALLES*, de la base de datos de *Scanntech*, y debe extraerse directamente de las compras realizadas. En dicha tabla, para cada compra, cada línea de detalle corresponde, en la mayoría de los casos, a un artículo. En los casos restantes la compra refiere a un rubro, que es una forma de clasificación general de los artículos. A los efectos de la construcción del DW, la unión del conjunto de todos los artículos con el de todos los rubros referidos en las compras conforma el conjunto total de productos.
Debido a que la codificación de los artículos y los rubros es responsabilidad de cada comercio, se vuelve una tarea difícil identificar cuándo dos artículos de diferentes comercios se están refiriendo al mismo producto.
En virtud de las dificultades encontradas se decidió:

- Crear un diccionario cuyo objetivo es la identificación de un mismo producto en diferentes comercios.
- Crear un prototipo de aplicación, en lenguaje Java, que permita realizar la categorización de productos.

Diccionario de Productos
El diccionario de productos está dividido en dos partes: una correspondiente a los artículos y otra a los rubros. Para su construcción, en cada instancia de actualización del DW se genera un archivo log con los artículos y los rubros, obtenidos de las compras, que no se encuentran ya en el diccionario. Este archivo es luego procesado por un procedimiento de ETL que actualiza el diccionario. A diferencia de lo que pasa con los artículos, los rubros no representan una cantidad importante, por lo que la parte del diccionario que les corresponde se construye en forma manual.
Para lograr identificar un mismo producto en comercios diferentes se utiliza el código de barras, el cual permite unificar los distintos artículos en un mismo producto por medio del diccionario. En la Tabla 5.3 se muestra un ejemplo de instancia, de un subconjunto de campos de la tabla *MOVIMIENTOS_DETALLES*, con artículos de diferentes comercios que representan el mismo producto. En este caso, el código de barras permite su identificación, y los artículos se agregan al diccionario con el mismo identificador de producto.

Código empresa	Código Artículo	Código Barra	Descripción
2246	000000000514586	7791293973180	SUAVE AC A/VERA
3047	000000000103362	7791293973180	ACONDICIONADOR
3082	000000000103802	7791293973180	SUAVE ACOND.PAL

Tabla 5-3: Identificación de artículos mediante su código de barras.
El campo *Código Artículo* identifica a un artículo dentro de un mismo comercio

Sin embargo, el procedimiento de identificación mediante el código de barras sólo se puede aplicar a una parte de los artículos. En primer lugar, los rubros, considerados como un producto más en el DW, no tienen código de barras. En la Tabla 5.4 se muestra un ejemplo en el que cuatro empresas utilizan códigos diferentes para identificar al mismo rubro. Para manejar estos casos se realiza una búsqueda en el diccionario de rubros utilizando la descripción de los mismos. El identificador en el diccionario de rubros es un identificador de un producto en la dimensión.

Código empresa	Código Artículo	Código Barra	Descripción Rubro	Código Rubro
2004	-	-	HIG. PERSONAL	26
2310	-	-	HIGIENE	23
2346	-	-	HIGIENE PERSONA	24
2372	-	-	ART.HIGIENE	4

Tabla 5-4: Identificación de rubros mediante su descripción.
El campo *Código Rubro* identifica a un rubro dentro de un mismo comercio

Por otra parte, existe otra clase de artículos cuyos códigos de barras son internos a la empresa. Estos artículos, generalmente, son los que se venden al peso. Sus códigos de barras comienzan con un 2, lo que indica que el código es interno. En estos códigos los comercios trasmiten información sobre la compra del artículo. Por ejemplo: para un mismo artículo el código puede variar según el peso y precio de la compra, ítems que quedan registrados dentro del código. Por lo tanto, en estos casos el código de barras no sirve para identificar al artículo ni dentro ni fuera del comercio. Para realizar la identificación sólo se tendrá en cuenta el código de la empresa y el código del artículo. La Tabla 5-5 muestra dos ejemplos de un mismo un artículo con diferente código de barras. En estos casos podemos

identificar la compra del artículo dentro del mismo comercio, pero no podemos identificar la compra de este mismo artículo en otro comercio. Para estos casos se realiza un post-procesamiento del diccionario basado en la descripción de los artículos, intentando identificar el mismo artículo en diferentes comercios.

Código empresa	Código Artículo	Código Barra	Descripción
2004	000000000002803	2002803010406	QUESO SEMIDURO
2004	000000000002803	2002803013001	QUESO SEMIDURO
2005	000000000000138	2001380027524	PAN DE OFERTA
2005	000000000000138	2001380029245	PAN DE OFERTA

Tabla 5-5: Artículos iguales con códigos de barras diferentes.

Otro caso a tener en cuenta es que existen productos diferentes que son considerados iguales para el DW. Éstos son, por ejemplo, las diferentes fragancias de un jabón de tocador. Si bien estos productos tienen distintos códigos de barra, no interesa diferenciarlos para este trabajo. Para reconocer estos casos nos fijamos en el código de empresa y código de artículo, aprovechando el hecho de que las empresas ya los consideran como el mismo artículo. La tabla 5-6 muestra un ejemplo en donde el mismo código de artículo para una empresa identifica productos con código de barra distintos.

Código empresa	Código Artículo	Código Barra	Descripción
2246	000000000514586	7791293973180	SUAVE AC A/VERA
2246	000000000514586	7791293973241	SUAVE AC A/VERA
2246	000000000514586	7791293999890	SUAVE AC A/VERA

Tabla 5-6: Diferentes versiones de un mismo artículo.

Luego del procesamiento del diccionario tenemos que los productos que son internos a una empresa no podemos cruzarlos con otra empresa porque no tenemos el código de barra. Debido a que son una cantidad importante buscamos una manera de identificar productos iguales en distintas empresas. Realizamos un post-procesamiento buscando descripciones de productos iguales, en caso de que sean el mismo producto unificamos el identificador del diccionario para que sean el mismo producto.

El diccionario creado con los datos mostrados en las tablas 5-3, 5-5 y 5-6 se muestra en la Tabla 5-7

Identificador Producto	Código empresa	Código Artículo	Código Barra	Descripción
1	2246	000000000514586	7791293973180	SUAVE AC A/VERA
1	3047	000000000103362	7791293973180	ACONDICIONADOR
1	3082	000000000103802	7791293973180	SUAVE ACOND.PAL
3	2004	000000000002803	2002803010406	QUESO SEMIDURO
3	2004	000000000002803	2002803013001	QUESO SEMIDURO
4	2005	000000000000138	2001380027524	PAN DE OFERTA
4	2005	000000000000138	2001380029245	PAN DE OFERTA
1	2246	000000000514586	7791293973241	SUAVE AC A/VERA
1	2246	000000000514586	7791293999890	SUAVE AC A/VERA

Tabla 5-7: Ejemplo de diccionario de productos

En la tabla 5-8 se muestra como quedaría el diccionario de rubros con los datos del ejemplo de la tabla 5-4

Identificador Producto	Descripción Rubro
2	HIG. PERSONAL
2	HIGIENE
2	HIGIENE PERSONA
2	ART.HIGIENE

Tabla 5-8: Diccionario de rubros

La Figura 5.7 muestra el flujo utilizado para la carga del diccionario.

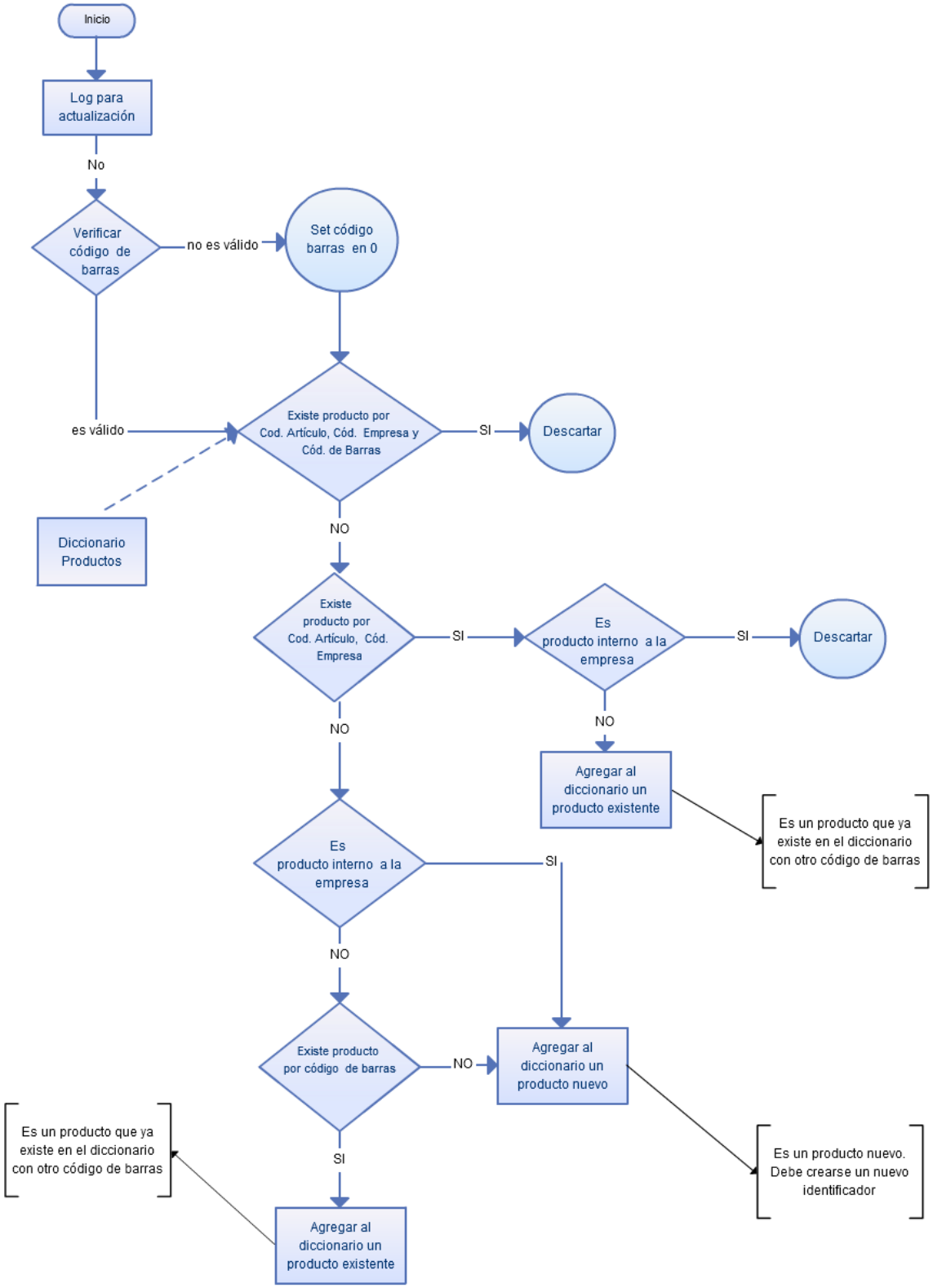


Figura 5-7: Flujo de carga: diccionario de productos.

Categorización de Productos

Los requerimientos incluyen un análisis de los productos por categoría. Debe entenderse por categoría a la agrupación de productos, según un criterio arbitrario, que permita a los usuarios realizar estudios mayormente de carácter nutricional. Algunos ejemplos podrían ser: Carnicería, Panadería, Arroz y Fideos. En los datos de origen no existe ningún tipo de agrupación, por lo tanto es necesario crearla como parte del desarrollo del DW.

Se decidió construir un prototipo de aplicación web con el objetivo de ayudar al usuario a clasificar productos mediantes sugerencias. A continuación se describen los pasos a seguir para realizar una categorización: en primer lugar, el usuario debe crear una nueva categoría. Luego debe indicar una lista inicial de productos que pertenecen a la categoría creada. Esta lista se utiliza como punto de partida para que el sistema comience a sugerir productos que integrarán la nueva categoría. El usuario podrá aceptar o rechazar las sugerencias.

Para generar las sugerencias, la aplicación se basa en la búsqueda de productos con descripciones similares a la de los productos ya existentes en la categoría.

El usuario selecciona las sugerencias que son correctas y descarta las restantes. Este paso se vuelve a ejecutar, tomando en cuenta los productos recientemente agregados, para generar las nuevas sugerencias. Los productos que fueron descartados no se vuelven a sugerir. El procedimiento se repite hasta que ya no hay más sugerencias. En ese momento, el usuario puede volver a buscar productos aún no categorizados y retomar el proceso, o bien dar por finalizada la categorización.

El cálculo de la similitud entre las descripciones de productos se realiza utilizando la transformación *Fuzzy LookUp,* de *Integration Services* (ver sección 5.2.2). El usuario puede elegir entre tres distintos umbrales de similitud para determinar si desea más resultados pero con menos probabilidad de acierto, o menos resultados pero más confiables.

Con esta aplicación se logra una automatización parcial de la categorización de productos. Como ejemplo: cuando se utiliza la opción más confiable, el sistema puede generar una lista extensa de sugerencias correctas y el usuario realizar la categorización con un solo click. Como contrapartida se tiene que hay productos que no aparecen en las sugerencias. Esto ocurre como consecuencia de que sus descripciones son elegidas por los comercios, y este hecho genera la existencia de una gran variedad de descripciones para un mismo producto. Dichos productos deben ser descubiertos y categorizados manualmente.

Los detalles del funcionamiento de la aplicación se encuentran en el Anexo 1.

Carga de la dimensión

La figura 5.8 muestra el flujo de carga para la dimensión *Producto*.

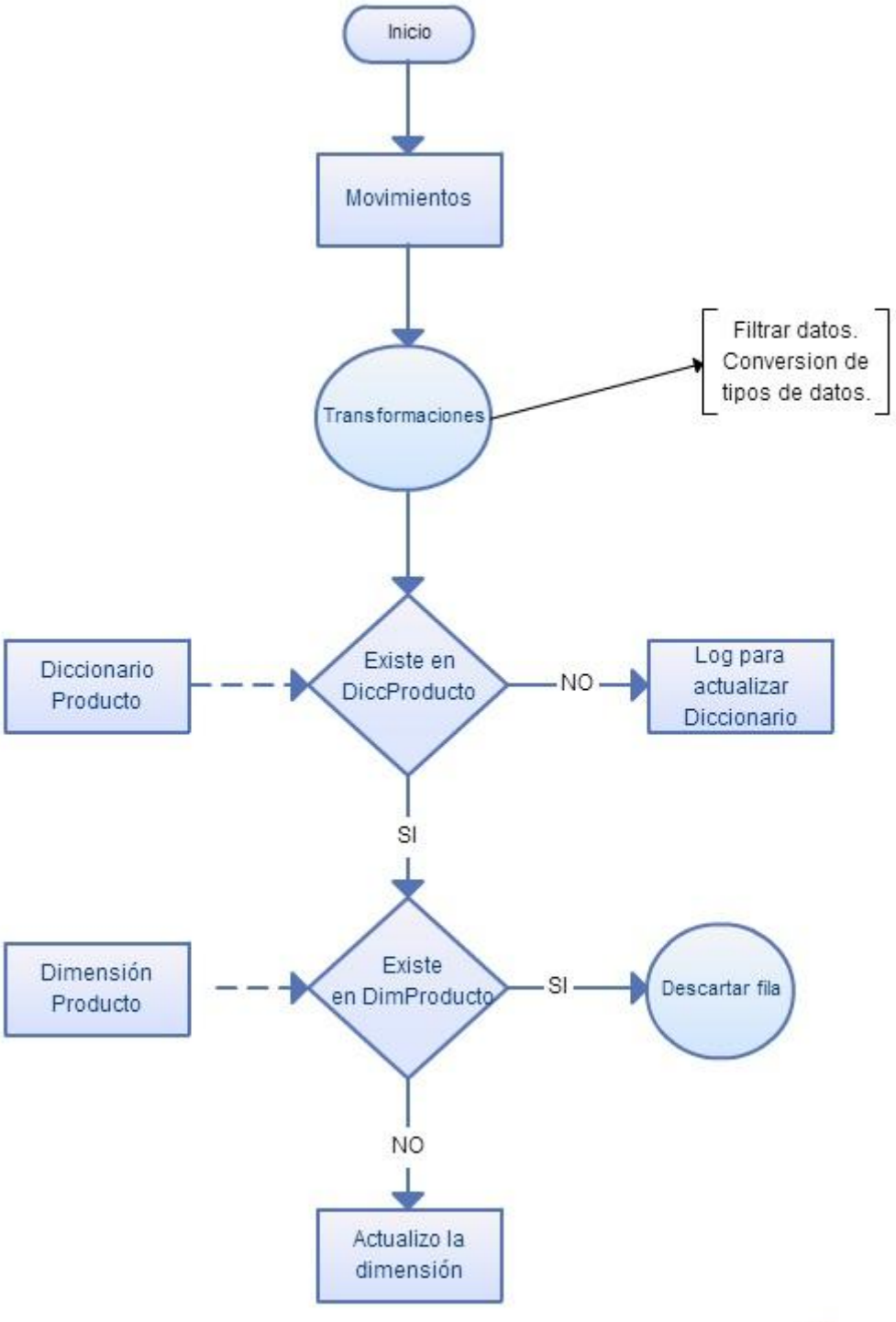


Figura 5-8: Flujo de carga: dimensión *Producto.*

5.3.3.3 Dimensión Persona

La dimensión *Persona* permite realizar el análisis según los datos socioeconómicos de la población beneficiaria. Se prevé que será utilizada por la mayoría de los programas, lo que la transforma en una de las dimensiones compartidas más importantes para el DW corporativo.

Problemas para la carga de datos.
Como se mencionó en la sección 4.2.5.1 (dimensión *Persona*), en el sistema informático utilizado para el registro de las compras realizadas con la TUS, el numero de cédula es ingresado de forma manual por el cajero al momento del pago. Analizando los datos fuente se puede deducir que el sistema no realiza restricciones sobre dicho ingreso, pues se pueden observar cédulas ingresadas con distintos formatos. A continuación se mencionan algunos ejemplos:

- Se encontraron cédulas con símbolos diferentes para separar el dígito verificador. Algunos de los símbolos encontrados son: punto, asterisco y guión (ejemplos: 1234567*8, 1234567.8, 1234567-8).

- Algunas cédulas tienen letras u otros símbolos ingresados después de su último dígito (ejemplos: 12345678*, 1234567a).

- No todas las cédulas contienen el dígito verificador y, entre las que sí lo tienen, existen algunas en los que dicho dígito no está separado de los dígitos restantes (ejemplos: 1234567, 12345678, 1234567.8).

Este problema se solucionaría con la existencia de un histórico de correspondencia entre el número de tarjeta y la cédula de la persona a quien le fue asignada, debido a que el número de tarjeta se registra en forma automática al momento de realizar la compra.
Otro dato a tener en cuenta es que las cédulas válidas pueden tener seis o siete dígitos más el digito verificador, el cual, en los datos de la TUS, puede estar presente o no. Por esto se deduce que el rango del largo de las cédulas con las que se trabaja está entre seis y ocho dígitos, considerando las cédulas con un largo fuera de dicho rango como erróneas.
Por los datos expuestos anteriormente se puede saber que:

- Las cédulas ingresadas con 6 dígitos son cédulas de seis dígitos sin el dígito verificador.

- Las cédulas ingresadas con 7 dígitos o bien son cédulas de seis dígitos con el dígito verificador o bien son cédulas de 7 dígitos sin el dígito verificador.

- Las cédulas ingresadas con 8 dígitos son cédulas de 7 dígitos ingresadas con el dígito verificador.

Durante el análisis de los datos se observan las siguientes cantidades aproximadas de cédulas según el largo de las mismas:

- Largo 6 – 40 cédulas.
- Largo 7 – 58000 cédulas.
- Largo 8 – 34000 cédulas.

Un último punto a tener en cuenta para el diseño de los procesos de carga es que, sobre la base de tarjeta, el único control que se puede realizar para saber si la cédula fue digitada correctamente es el del dígito verificador. Este control sólo puede ser aplicado a las cédulas ingresadas con ocho dígitos, pues éstas son las únicas en las que se puede confirmar que se ingresó el dígito verificador.

Solución implementada

La cantidad de personas con cédulas de 6 dígitos (menores a un millón) es muy poca en relación a las personas con cédulas de 7 dígitos (menos del 0.1%), por lo tanto, se considera que las cédulas ingresadas con 7 dígitos son cédulas ingresadas sin el digito verificador, y no cédulas de 6 dígitos con dígito verificador.

Para obtener los datos de las personas utilizamos dos orígenes de datos: uno consiste en la base de datos de AFAM, brindada por BPS, y el otro en los archivos de texto con los datos con los que MIDES alimenta a SIIAS. Es de destacar que estas bases sí tienen la cédula de las personas con un formato único: cédula con el dígito verificador sin puntos ni guiones, por ejemplo, la CI 1234567-8 se registra como 12345678.

El primer paso del proceso de limpieza consiste en eliminar, de los números de cédula, todos los símbolos que no sean un dígito. El siguiente paso es determinar si las cédulas contienen, o no, su dígito verificador.

La carga de las personas se divide en tres etapas según el largo de la cedula ingresada (8 ,7 y 6 dígitos). En una primera etapa se cargan las cedulas de 8 dígitos, pues estas son las únicas que se pueden verificar. En caso de fallo en la verificación, la persona se carga en la dimensión sólo con los datos de su cédula y tarjeta (sin sus restantes datos básicos). De esta manera se expresa que la información correspondiente a la persona no es confiable, y si en una próxima carga existe una persona con la misma tarjeta y una cédula que verifique, se cambia la cédula anterior por la nueva y se agregan los datos básicos.

En una segunda etapa se cargan las cédulas ingresadas con 7 dígitos. Dado que la probabilidad de que esta sea una cedula de largo 7 sin el digito verificador es mayor que el 99,9% (la probabilidad de que sea una cédula de largo 6 con el digito verificador es menor

que 0,1%) buscamos esa cédula en las fuentes de datos de las personas (AFAM y SIIAS), en caso de encontrarlos se agrega a la dimensión con los datos obtenidos en esa base.
En la última etapa se cargan las personas con las cédulas ingresadas con 6 dígitos.
Para esta carga no sólo se relaciona la persona por la cédula sino también por la tarjeta, por lo que si una tarjeta ya está en la dimensión, y hay en los datos a cargar información con esa misma tarjeta y otro número de cédula, se envía a un log de conflicto para que sea analizado y resuelto manualmente.

Dimensión demográfica

Problemas con la carga de datos:
La carga de los datos demográficos se realiza desde distintas fuentes de datos, las cuales pueden variar entre dos cargas consecutivas (cargas de distintos meses). Además se necesita una solución que permita a un usuario corregir o modificar manualmente datos demográficos de determinadas personas, validando la veracidad de los mismos. Otra necesidad es la de poder integrar distintas fuentes de datos: por ejemplo si en un mes se realiza un relevamiento de un subconjunto de beneficiarios, se quiere que los datos demográficos de esos beneficiaros se actualicen con la información relevada y no con las fuentes de datos que se trabajan habitualmente.

Solución implementada:
Para lograr una solución que se adapte a los problemas mencionados se crea la tabla *DatosDemograficos* en el ODS, la cual contiene la información demográfica actual de las personas y funciona como interfaz entre las distintas fuentes de los datos demográficos y el DW. Con esta solución los datos demográficos de una persona en el DW pueden ser validados, verificados y modificados por el usuario del DW, simplemente actualizando la información de las personas en la tabla creada en el ODS sin alterar los procesos de carga automatizados. Esos cambios se introducen en la siguiente actualización del DW.
El proceso de carga automático busca los datos demográficos de las personas en la tabla del ODS cargada previamente y, en caso de que alguna persona haya modificado alguno de sus datos demográficos, se actualiza el valor de la *Dimensión Demográfica* actual correspondiente a esa persona.
Para este proyecto esta tabla se carga solamente con los datos de AFAM, ya que la fuente de datos de SIIAS utilizada no contiene información demográfica.

En resumen, con esta solución se logra lo siguiente:

- Los datos demográficos se pueden obtener de distintas fuentes sin modificar los procesos de carga automáticos.

- El ODS permite al usuario modificar la información demográfica de determinadas personas para actualizar en la próxima carga.

- Si se quiere impactar un cambio en los datos demográficos de una persona sólo hay que ingresar esa información en el ODS.

En la Figura 5.9 muestra como es el proceso de actualización de esta dimensión

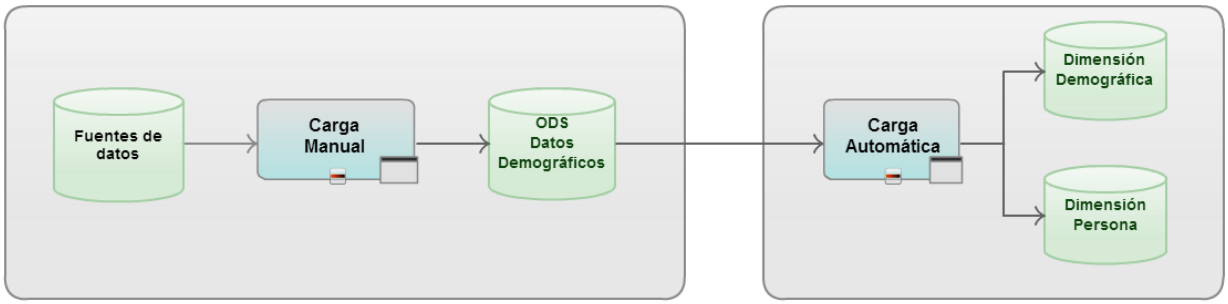


Figura 5.9: Proceso de actualización de la Dimensión Demográfica

Las Figuras 5.10 y 5.11 muestran los flujos de carga de la Dimensión Persona y Dimensión Demográfica respectivamente.

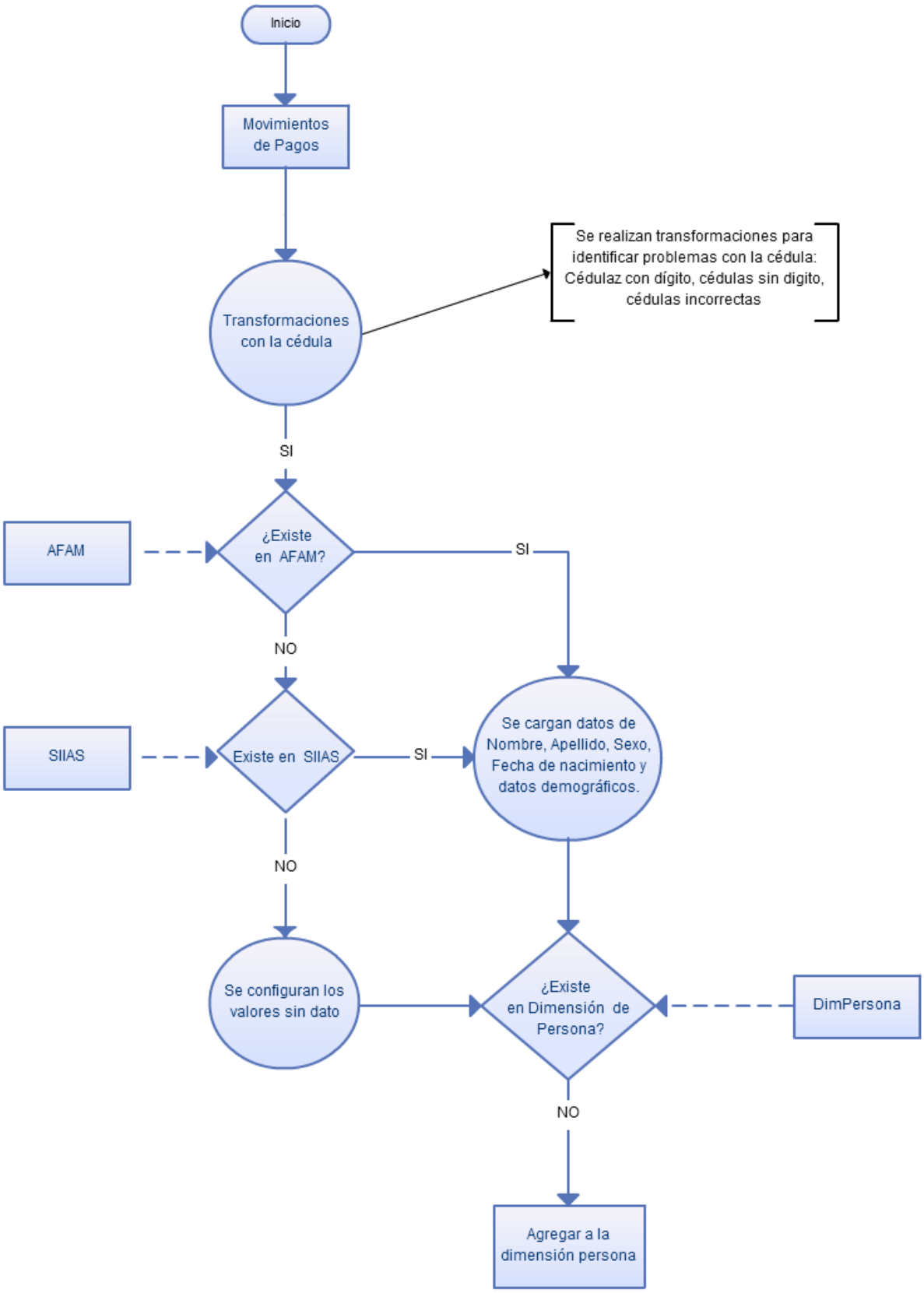


Figura 5-10: Flujo de carga: dimensión *Persona.*

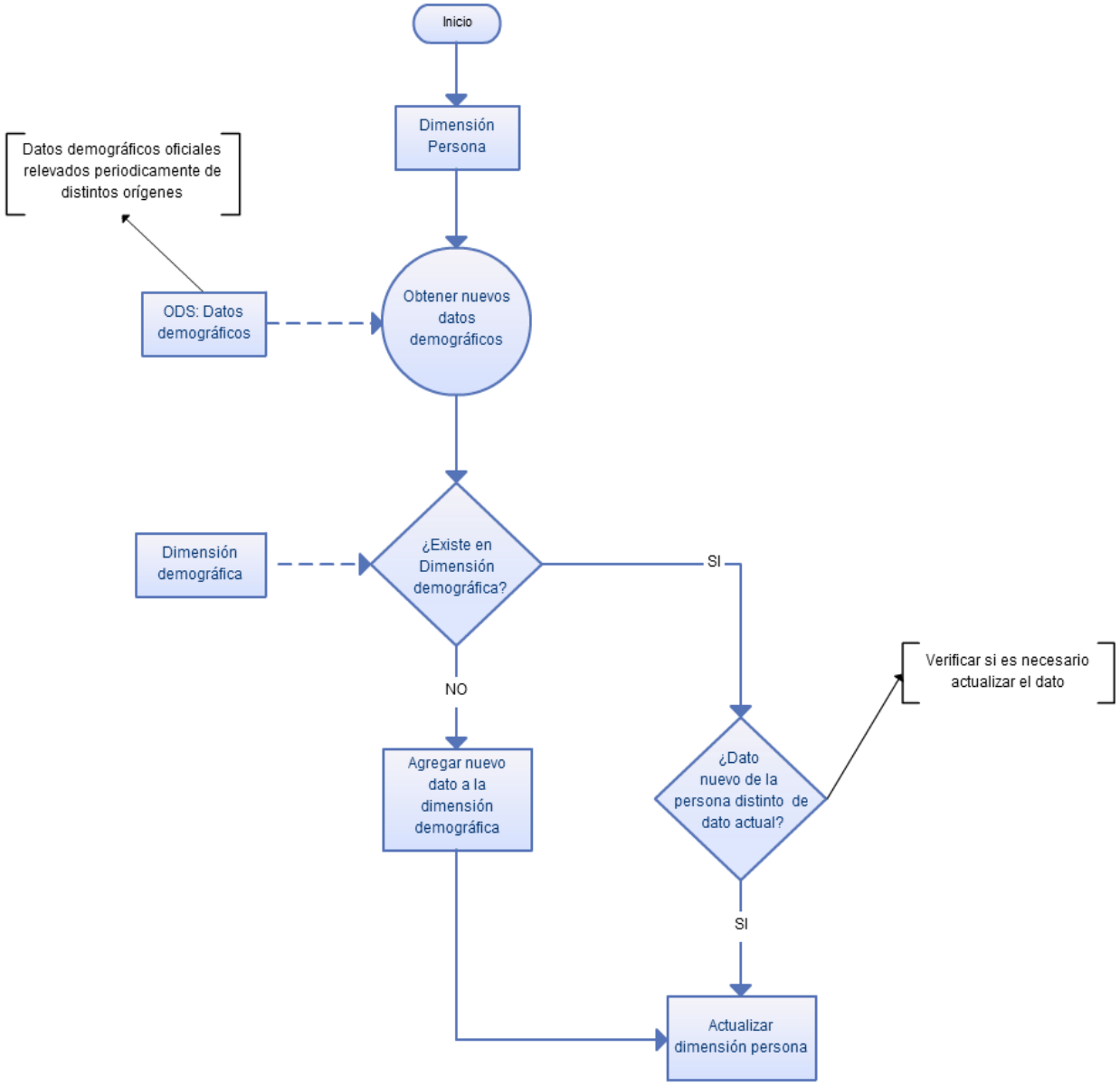


Figura 5-11: Flujo de carga: dimensión *Demográfica.*

5.3.3.4 Tabla de hechos: Fact Productos

La Figura 5.12 muestra el flujo de carga para la tabla de hechos *Fact Productos*.

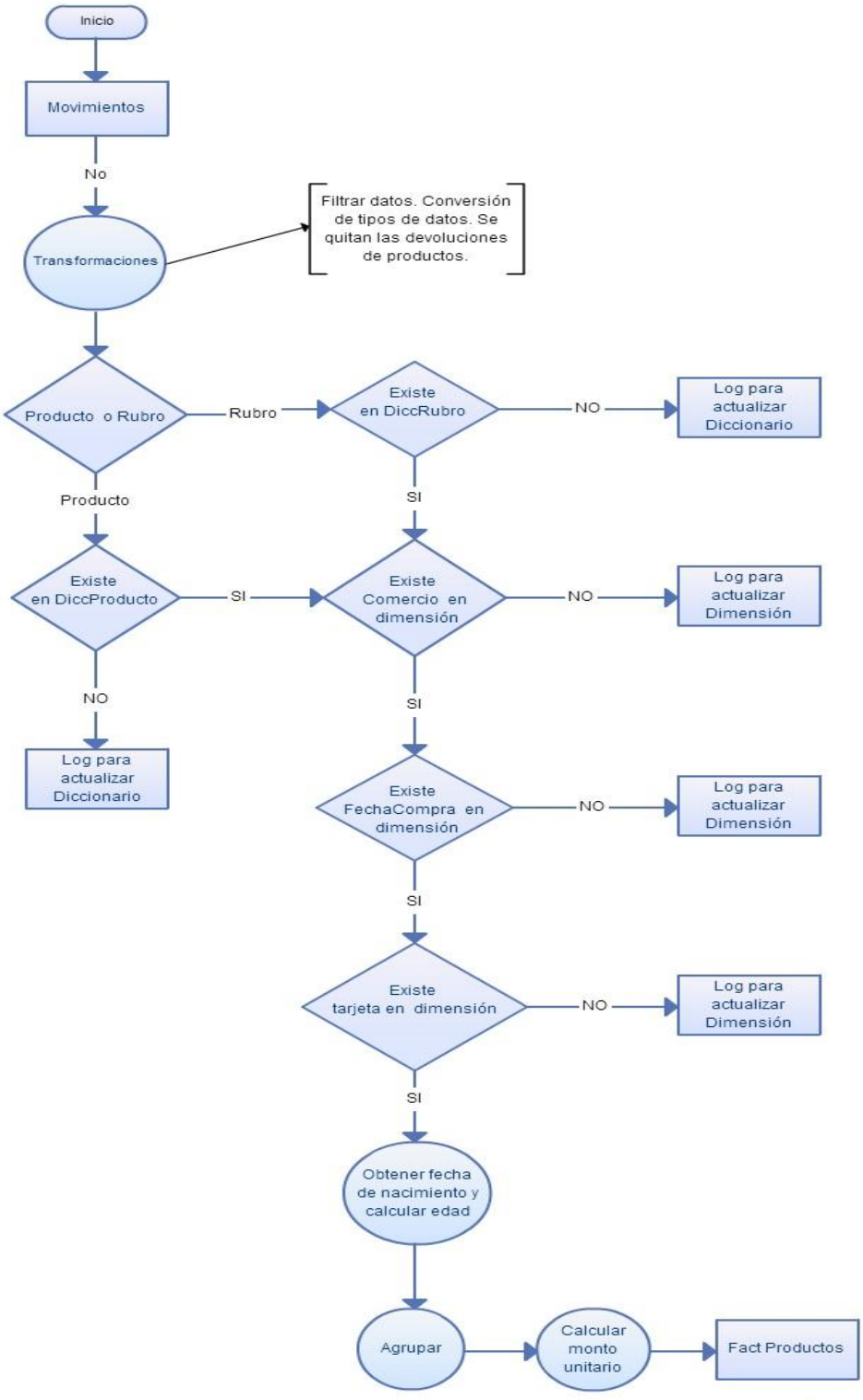


Figura 5-12: Flujo de carga: Tabla de hechos *Fact Productos*

5.3.3.5 Tabla de hechos: Fact Gastos

La figura 5.13 muestra el flujo de carga para la tabla de hechos *Fact Gastos*.

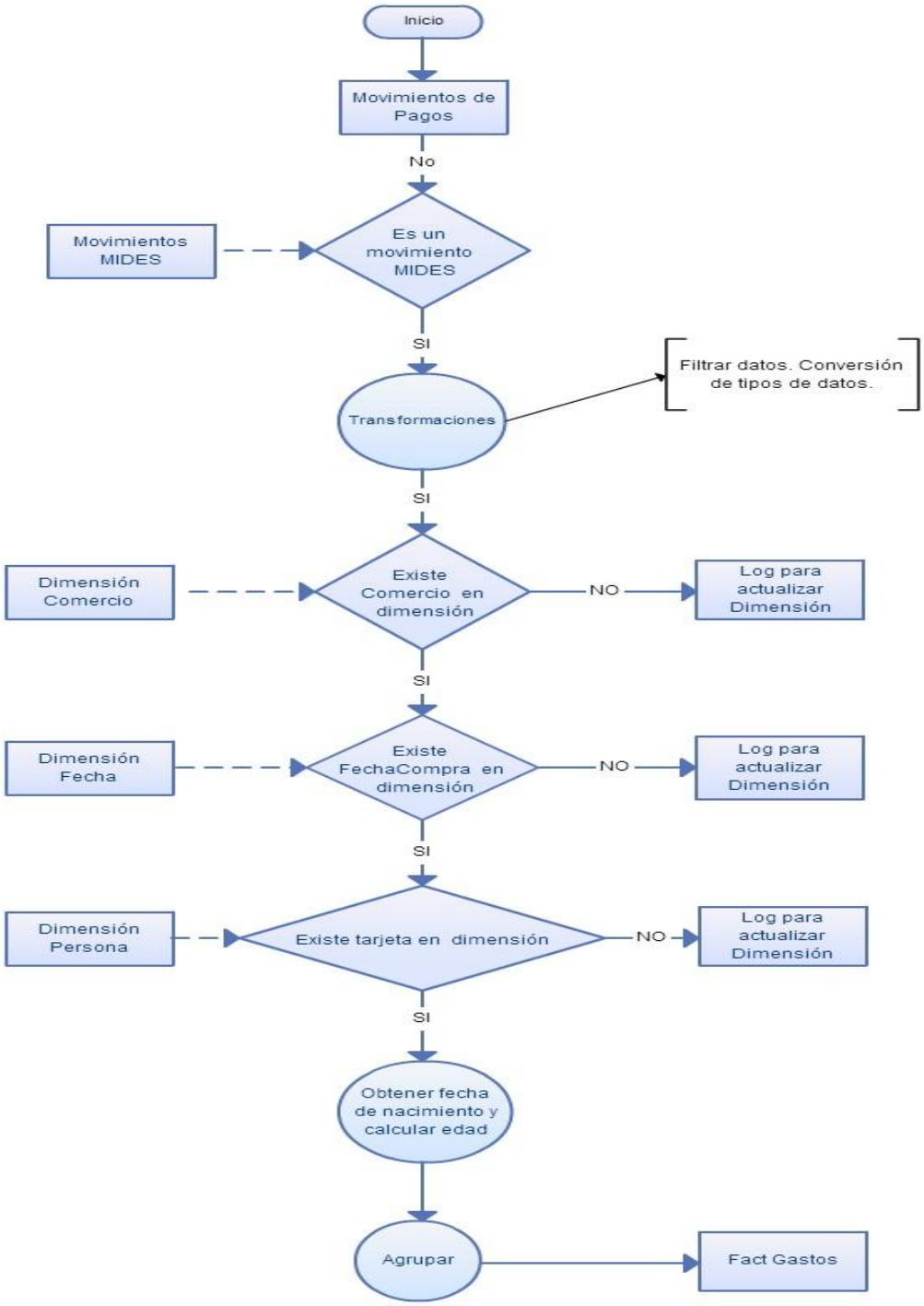


Figura 5-13: Flujo de carga: Tabla de hechos *Fact Gastos.*

6 Implementación

Este capítulo describe aspectos de la implementación de la solución. Se describen los datos utilizados, aspectos de la verificación y validación con los usuarios y una reseña de las herramientas utilizadas. También se adjuntan algunas imágenes de los resultados obtenidos.

6.1 Datos

A continuación se describen brevemente los datos utilizados para la implementación del DW.

Los datos del programa *Uruguay Integra* fueron proporcionados por la Dirección Nacional de Evaluación y Monitoreo de MIDES. Éstos constan de dos planillas Excel para 2010 y tres para 2009. Para el año 2010 se tiene una planilla con datos de inscriptos y grupos con 670 filas y una con datos de recursos humanos con 113 filas. Para el año 2009 se tiene una planilla con datos de inscriptos con 1916 filas, datos de grupos con 306 filas y datos de recursos humanos con 342 filas. Todos estos datos cargados al DW resultaron aproximadamente unas 3000 filas divididas en 5 tablas de hechos.
Para el caso de *Tarjeta Uruguay Social* los datos fueron provistos por la empresa *Scanntech*. Los datos corresponden al mes de enero de 2010 y sus cuatro tablas principales contienen las cantidades de filas glosadas en la tabla 6-1.

TABLA	CANTIDAD DE FILAS
MOVIMIENTOS_DE_PAGOS	273.472
MOVIMIENTOS	226.572
MOVIMIENTOS_DETALLES	1.970.419
MI_MOVIMIENTOS_PAGOS	278.505

Tabla 6-1: Cantidad de filas de las tablas principales de *TUS*.

Los datos cargados al DW resultaron en aproximadamente 1.800.000 filas divididas en 2 tablas de hechos.

6.2 Verificación y validación

Verificación

La verificación se realizó mediante consultas SQL específicas con el fin de comparar la información de las fuentes con los datos cargados en el DW. Este método permitió encontrar y solucionar errores en los procesos de carga de los datos. Los resultados de estas consultas también se utilizaron para verificar que la información de los cubos se cargó en forma correcta.

Para las tablas de hechos se calcularon los totales de las medidas en las fuentes de datos y se compararon estos valores con los valores cargados en el DW. Luego se seleccionaron valores arbitrarios de las dimensiones del DW para filtrar información, y se comparó el resultado de las medidas con la información original en las fuentes de datos. Dada la complejidad de la carga de la dimensión *Producto*, mencionada en la sección 5.3.3.2 (Dimensión Producto), se realizaron pruebas tomando productos arbitrariamente y verificando que no se hubieran unificado productos diferentes.

Para el caso del programa TUS, los usuarios cotejaron la información de los cubos con el conocimiento funcional que tienen del programa, obteniéndose una respuesta satisfactoria de parte de ellos.

Validación

La validación de los cubos construidos para el programa TUS se realizó en conjunto con los usuarios, quienes utilizaron la herramienta para generar sus propias consultas. Obtuvimos una respuesta satisfactoria de los usuarios manifestando el interés de agregar la información más reciente al DW para poder analizar estos datos periódicamente.

En el caso del programa UI no pudo realizarse la validación de los cubos construidos. No se pudieron concretar instancias de validación con los responsables funcionales del programa.

6.3 Resultados

A continuación se muestran algunos Snapshots de cómo se analizó la información en las distintas herramientas.

En la Figura 6-1 se puede ver una gráfica de barras de Microsoft Excel en la cual se muestra la cantidad de beneficiarios distribuidos por los diferentes departamentos.

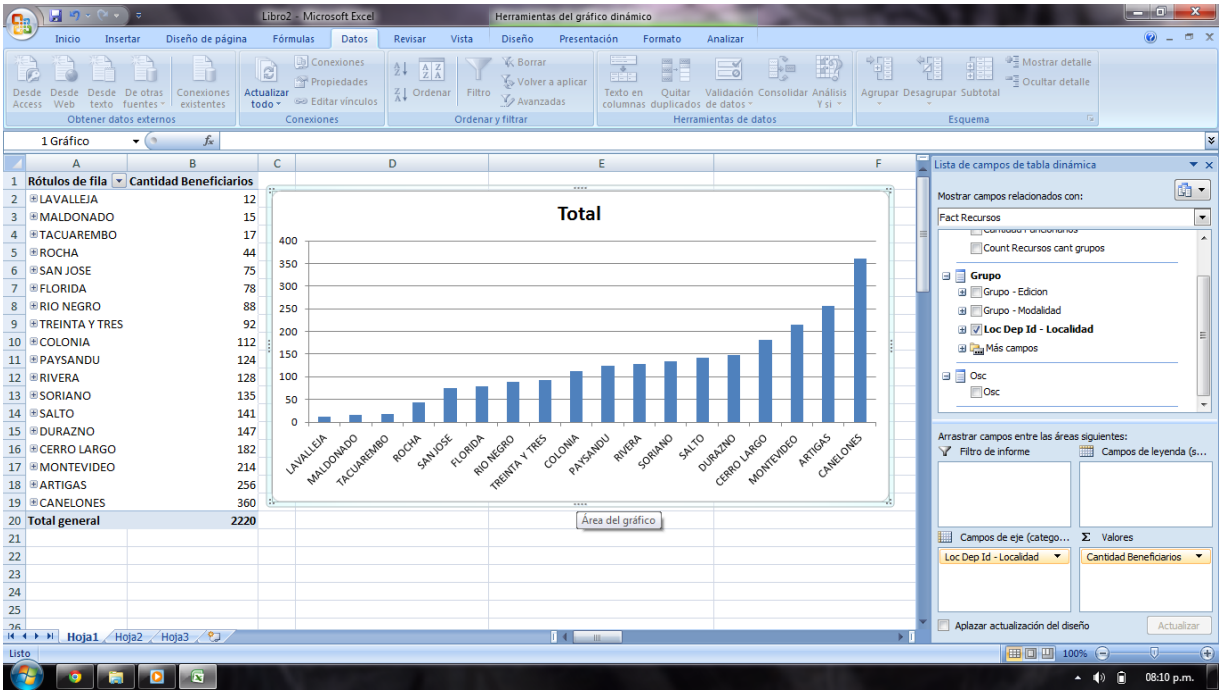


Figura 6-1: Microsoft Excel – Uruguay Integra

La Figura 6-2 muestra una gráfica de líneas de Microsoft Excel en la cual se observa el gasto por día para el mes de enero de 2010 en Montevideo.

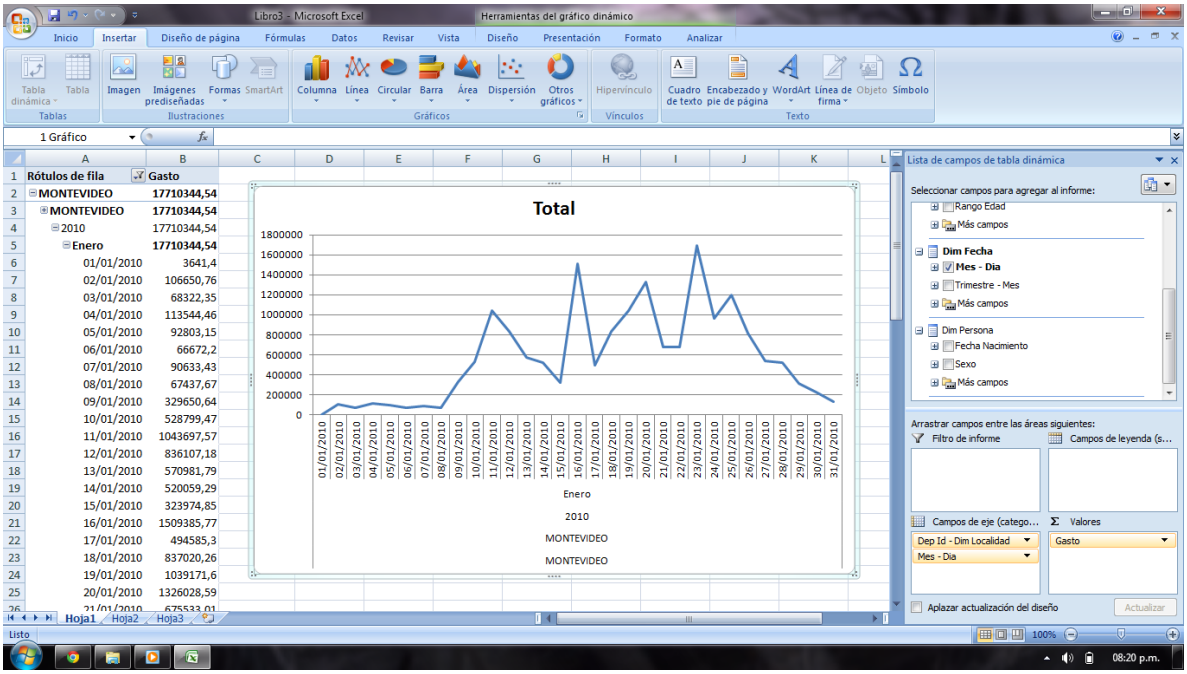


Figura 6-2: Microsoft Excel – Tarjeta Uruguay Social

La Figura 6-3 muestra un análisis realizado en la herramienta O3, el cual contiene una tabla con la información de la productividad de cada OSC. Esta productividad viene dada por la relación entre la cantidad de beneficiaros con la cantidad de grupos que gestiona dicha OSC. En el mismo análisis se puede ver una gráfica de barras con la cantidad de beneficiarios por OSC

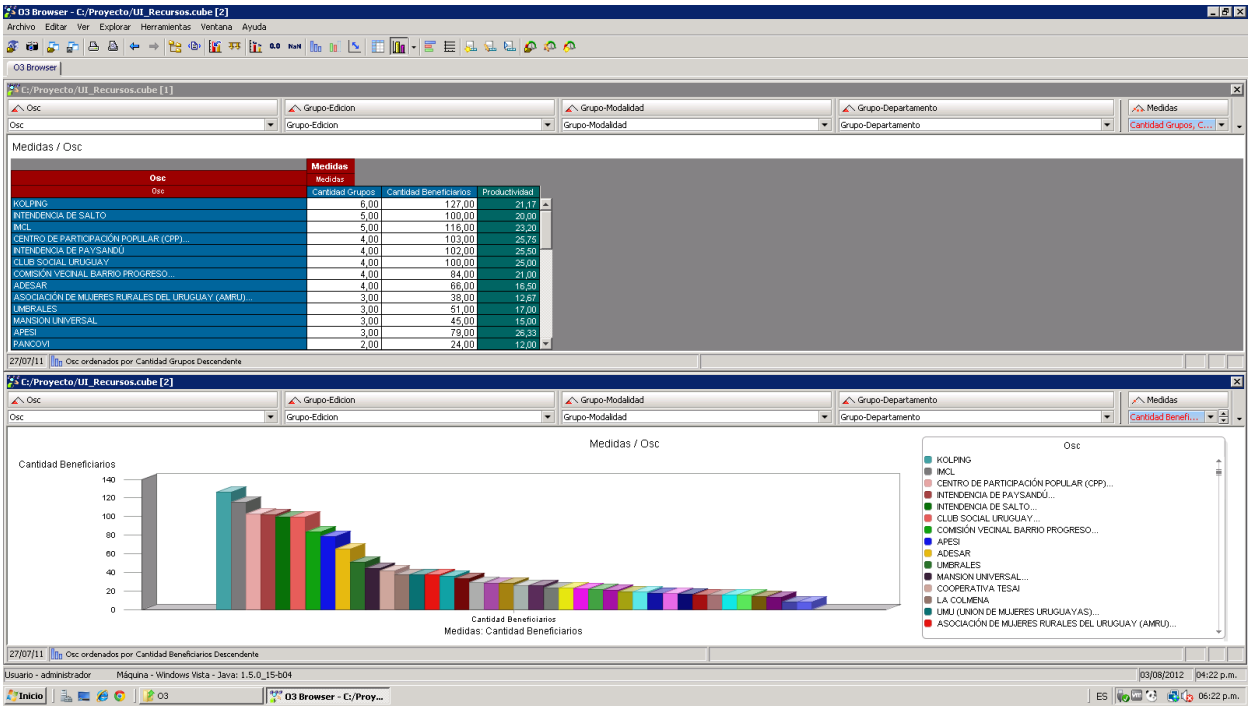


Figura 6-3: O3 –Uruguay Integra

En la Figura 6-4 se puede observar una gráfica de líneas, realizada en O3, en la cual se analiza la distribución del gasto por día en el mes de enero de 2010 según rangos de edad.

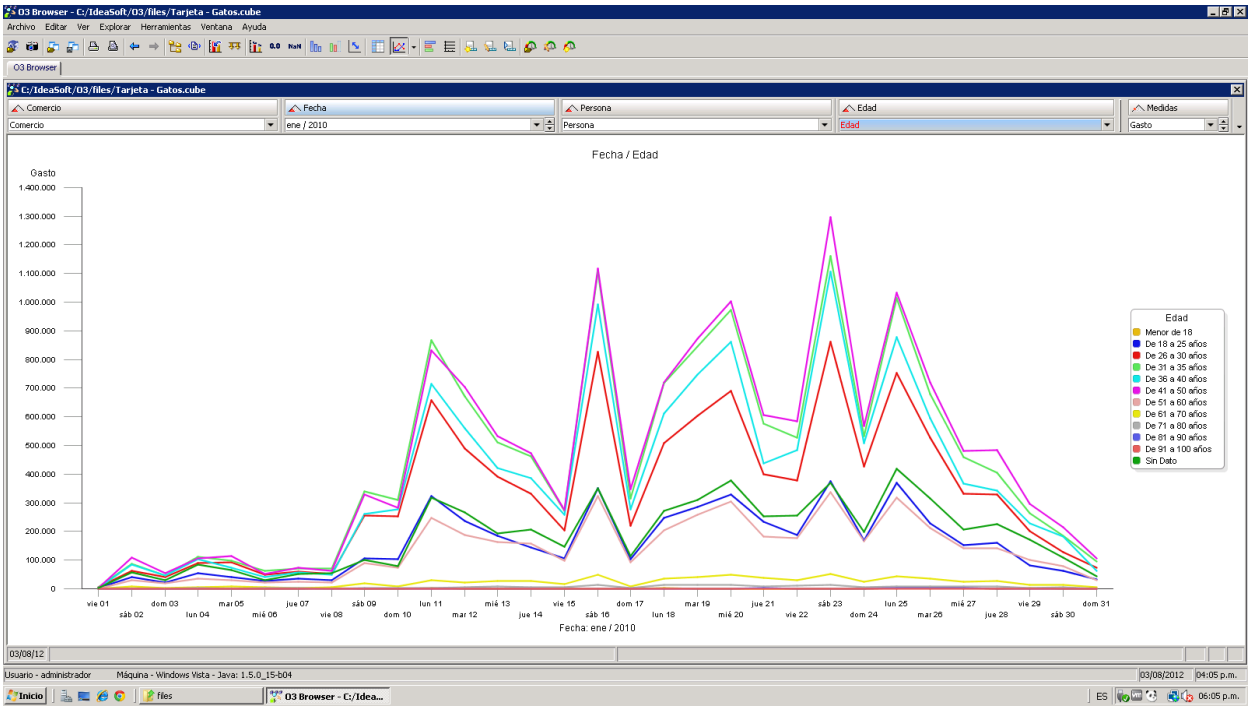


Figura 6-4: O3 – Tarjeta Uruguay Social

La Figura 6-5 muestra un análisis, en la herramienta Analysis Services, en donde se puede observar una tabla con la información de la cantidad de beneficiarios de UI analizada según las dimensiones *Sexo* y *Nivel educativo*.
Esta herramienta es utilizada durante el desarrollo para verificar los resultados pero no es utilizada por los usuarios para analizar la información.

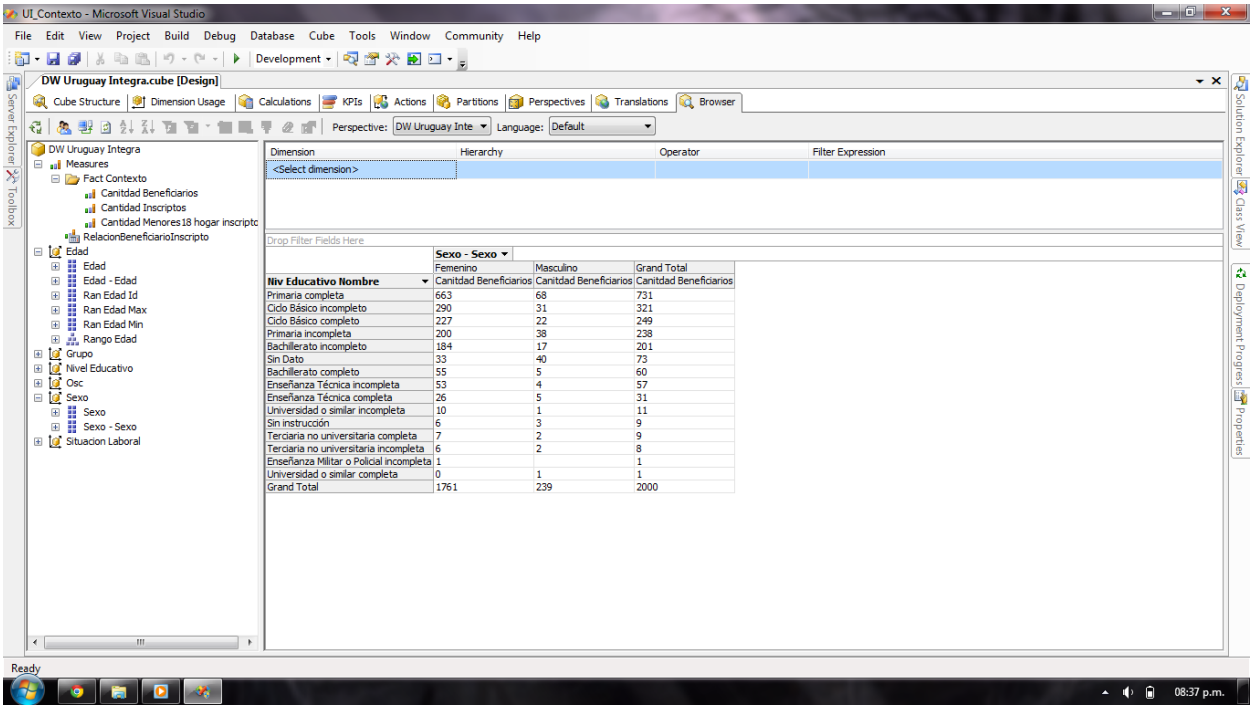


Figura 6-5: Analysis Services – Uruguay Integra

La Figura 6-6 muestra una tabla, realizada en *Analysis Services*, que muestra el gasto agrupado por categorías de productos para el autoservice YULIVAN de la localidad de Bella Unión, en el departamento de Artigas.

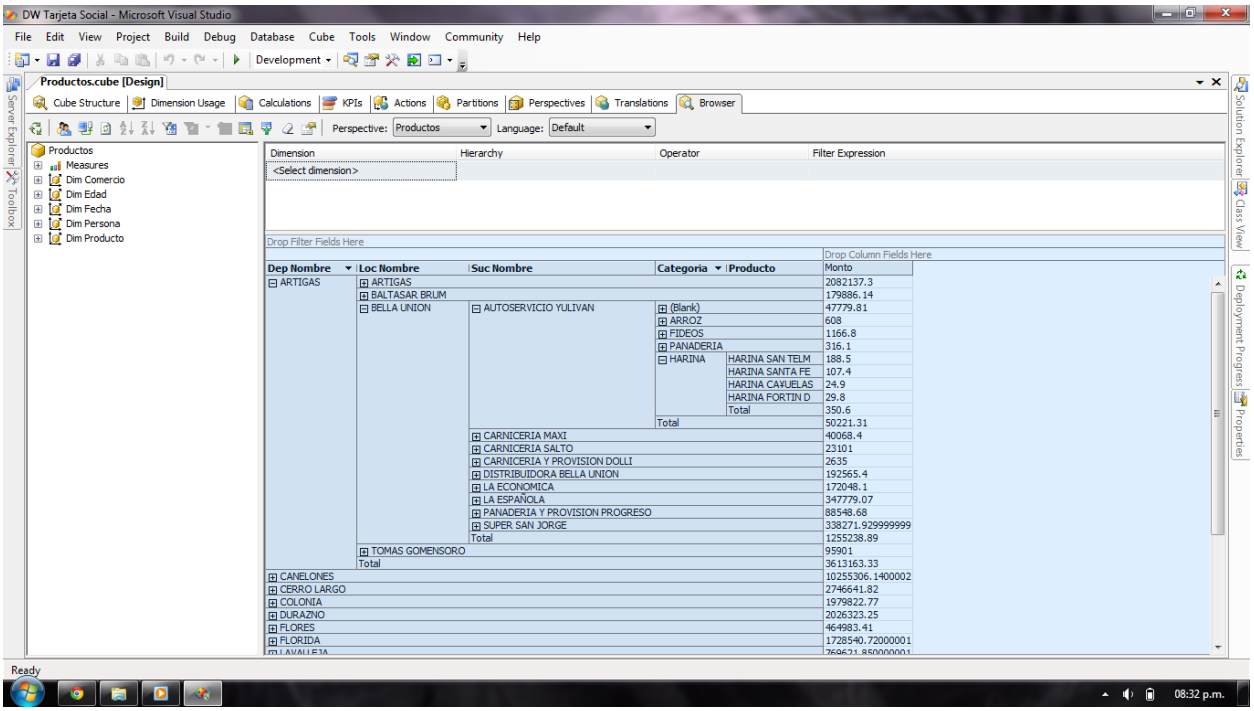


Figura 6-6: Analysis Services – Tarjeta Uruguay Social

6.4 Herramientas

En esta sección se describen las herramientas utilizadas en el proceso de construcción del prototipo y para el análisis de la información generada.

6.4.1 ETL

A continuación se realiza una breve reseña de las herramientas utilizadas para implementar los procesos ETL.

Microsoft Integration Services

Integration Services es una plataforma para la creación de soluciones para la transformación e integración de datos. Provee transformaciones que permiten por ejemplo descarga de archivos, envío de mensajes de correo electrónico, conexión con bases de datos, limpieza y minería de datos, etc. Permite extraer y transformar datos de diversos orígenes como archivos XML, archivos planos y orígenes de datos relacionales y luego

cargar los mismos en uno o varios destinos. Posee una potente interfaz gráfica que permite diseñar los flujos de transformaciones sin escribir código. Se arrastran transformaciones predefinidas mediante el sistema *drag and drop* y se construye un flujo de trabajo uniendo las mismas por medio de flechas. En caso de ser necesario se pueden escribir transformaciones con código *.NET* para diseñar una solución particular a un problema específico. *Integration Services* se adecuó satisfactoriamente a las necesidades del proyecto, no encontrándose mayores dificultades durante la implementación.

Pentaho Data Integration

Como ya se mencionó en el capítulo 3, referente a la arquitectura de la solución, uno de los requerimientos no funcionales consiste en tener alternativas a la herramienta de *Microsoft* para la carga de datos al Data Warehouse. Siguiendo la sugerencia de los consultores que elaboraron el *Plan Director de Sistemas Informáticos* [1], se optó por utilizar *Pentaho Data Integration*. Esta herramienta tiene características similares a la de *Microsoft*, provee una interfaz gráfica *drag and drop* que permite crear un flujo de datos al que se le aplican transformaciones. Posee una configuración muy simple y sólo necesita tener instalado *Java* en la máquina que se desee utilizar. Posee una variedad de transformaciones predefinidas que permiten realizar operaciones sobre Strings, conexiones a bases de datos, trabajar con archivos de texto plano, archivos Excel, etc. Como transformación destacada encontramos la *Fuzzy match*, a diferencia de la implementación de *Integration Services*, ésta permite configurar el algoritmo por el cual se quiere realizar el cálculo de la distancia. Se pueden elegir más de diez algoritmos para calcular la distancia entre dos cadenas de texto, como por ejemplo la distancia de Levenshtein, SoundEx, Jaro Winkler, etc. Al igual que *Integration Servicies*, en caso de no encontrar una transformación adecuada para cierto problema particular, también permite definir un paso de script en donde se puede especificar mediante código *Java* la solución al problema.

Se encontraron ciertas dificultades con la herramienta ocasionadas por la mala administración de la memoria, además de la existencia de bugs que, en algunos casos, llevan un largo tiempo de reportados y no han sido solucionados.

Evaluación

Con el fin de realizar la comparación de estas dos herramientas, ambas fueron utilizadas en la implementación de los procesos ETL del programa *Uruguay Integra.* Durante la prueba se observó una mayor eficiencia para realizar las cargas por parte de *Integration Services*. En cuanto al diseño de un proceso de ETL no se observaron mayores diferencias, la curva de aprendizaje para la implementación fue similar para ambas herramientas.

Ante una eventual migración a software libre, recomendamos *Pentaho Data Integration* como una buena opción alternativa.

6.4.2 Definición de cubos y análisis OLAP

Microsoft Analysis Services / Excel

Analysis Services ofrece funciones de procesamiento analítico en línea (OLAP). Permite la creación y administración de estructuras multidimensionales. Para la mayoría de los casos *Analysis Services* reconoce automáticamente, a partir del modelo de base de datos, cuales tablas son dimensiones y cuales tablas de hechos. Esto permite acelerar considerablemente el tiempo de creación de un cubo. La interfaz de la herramienta es clara y permite configurar una gran cantidad de parámetros; por ejemplo, se puede elegir la manera que en que se quiere persistir el modelo entre las opciones MOLAP, ROLAP u HOLAP. Para algunos casos particulares fue necesaria una implementación en archivos MDX. Un punto bajo de la herramienta es la presentación de la información al usuario. De todas maneras esto no fue un problema debido a que permite que la información sea leída desde un archivo de Excel como una tabla dinámica. Los usuarios están más familiarizados con el uso de Excel y les resulta fácil aprender a manejar tablas dinámicas y elaborar gráficas y consultas de acuerdos a sus necesidades. No encontramos limitaciones a la hora de implementar los cubos para el proyecto.

Ideasoft O3

O3 es una herramienta que permite diseñar y crear estructuras multidimensionales, y luego analizar los resultados mediante un portal en donde se pueden generar gráficas dinámicamente. *O3 Designer*, componente para el diseño de cubos, permite la creación de dimensiones y tablas de hechos por medio de consultas que indican el origen de los datos.

La interfaz de usuario resulta menos intuitiva que la de *Analysis Services,* y los errores que se presentan no tienen una clara descripción. A esto se le suma que la información disponible en internet es escasa.

Una limitación encontrada en el *O3 Designer* es que no es posible crear múltiples jerarquías de agregación para una dimensión. Para solucionar esto es necesario crear una nueva dimensión para cada jerarquía.

Las funciones de agregación deben seleccionarse desde un conjunto predefinido de operaciones, no permitiendo la definición de nuevas funciones.

Para la presentación de la información al usuario utilizamos el componente *O3 Portal*, el cual es el componente más destacable de *O3*. Es un componente WEB que, utilizando el sistema *drag and drop,* permite elegir las medidas y dimensiones actualizando dinámicamente las gráficas en donde se presenta la información. Con este componente se puede presentar la información de una manera clara y visualmente atractiva. *O3 Portal* también permite compartir información mediante vistas generadas por los propios usuarios.

No se encontraron mayores dificultades con la herramienta y permitió cumplir satisfactoriamente con los requerimientos.

Evaluación

Se trata de dos herramientas con marcadas diferencias que, sin embargo, pueden complementarse de buena forma. La herramienta *Analysis Services* es más potente que *O3* en lo que respecta a la definición e implementación de los cubos, mientras que la fortaleza de *O3* radica en la presentación de la información: es una herramienta fuertemente enfocada al usuario. A pesar de esto se observó que, a la hora de interactuar con la herramienta, los usuarios prefieren utilizar Excel.

7 Conclusiones y trabajo a futuro

Resultados alcanzados

- Se logró construir una solución diseñada con una arquitectura en capas, la cual facilita la incorporación de nuevos programas sociales y nuevas fuentes de datos. Esta arquitectura permite, además, lograr independencia de las herramientas utilizadas.

- Se cargó al DW construido la información de los programas sociales Uruguay Integra y Tarjeta Uruguay Social.

- Se implementaron prototipos de los cubos de ambos programas sociales en las herramientas *Analysis Services* y *O3*.

- Se definieron e implementaron los procesos de carga y actualización automáticos del Data Warehouse.

- Se implantó la solución en el ambiente de testing brindado por el MIDES.

Dificultades encontradas

- Una de las mayores dificultades encontradas durante el proceso obedeció al difícil acceso a las fuentes de datos de TUS. Como fue mencionado en la sección 4.2.2, el sistema de información correspondiente a TUS está tercerizado. Si bien la empresa que lo administra expresó su voluntad para brindar los datos necesarios, finalmente la interacción fue muy lenta, demorándose varios meses, además de que sólo se pudo obtener los datos correspondientes a un mes de movimientos.

- El hecho de no conocer el sistema de TUS, y de no contar con ninguna documentación ni un referente a quien consultar, obligó a descifrar su funcionamiento a través del estudio de los datos brindados por la empresa administradora de los mismos. Esto significó una gran inversión de tiempo en detalles que no hubieran representado mayor dificultad en caso de tener un mínimo conocimiento del sistema. A manera de ejemplo se puede mencionar el caso de las devoluciones de productos: no saber que estaban registradas indujo a cargas de información errónea, hecho advertido recién en la etapa de verificación.

- La mala calidad de los datos de UI obligó a destinar gran parte del tiempo de desarrollo en la limpieza de los mismos.

- Si bien en la organización existe consciencia sobre la necesidad de consolidar información en un DW corporativo, la asignación de recursos destinados al trabajo requerido para su construcción y mantenimiento no se encuentra actualmente entre las prioridades del ministerio. Los usuarios están acostumbrados a la manipulación artesanal de las fuentes de datos para el análisis de la información, y se sienten cómodos con las herramientas que utilizan actualmente.
  Los intentos por imponer al DW como solución vienen dados por esfuerzos individuales y no por una política clara e impulsada desde los altos mandos.

Conclusiones finales

La implementación para el programa UI no colmó nuestras expectativas teniendo en cuenta la relación entre el esfuerzo de trabajo que representó la automatización de los procesos, de limpieza y carga, y los resultados obtenidos. El programa contiene una cantidad de datos relativamente pequeña y con muy mala calidad. La limpieza de los mismos podría haber resultado más rápida si se hubiera realizado en forma manual en vez de implementar procesos de limpieza automáticos.
Algo muy distinto ocurre con el programa TUS. La cantidad de datos es muy grande y la información no se encuentra disponible en las bases de datos del Ministerio, por lo que no se podría realizar un análisis eficiente de la información sin contar con una solución como la implementada en este proyecto. Para TUS se logró construir una solución que permite analizar todas las compras realizadas por los beneficiarios de manera rápida y sencilla, con distintas alternativas de visualización según preferencias del usuario. Agregando valor a la información disponible en las bases de Tarjeta gracias a la integración con distintas fuentes de datos que aportaron variables socioeconómicas para el análisis.

Finalmente se concluye que la realización de este proyecto representó un primer paso hacia la construcción de un DW corporativo en el MIDES. Quedaron implementados los procesos y las estructuras que facilitarán en el futuro la incorporación de nuevos programas sociales.

Trabajo a futuro

- Continuar agregando los distintos programas sociales que implementa el Ministerio. También sería deseable que se enriqueciera la Dimensión *Persona* con nuevos datos socioeconómicos, lo que son muy importantes para el análisis que se realiza en el MIDES.

- Incluir indicadores de gestión y nuevos reportes de alto nivel para brindar información a los directores de los distintos programas sociales.

- Mejorar la aplicación de categorización de productos para poder automatizar la mayor parte posible de este proceso.

8 Referencias

[1] Gabriela Berch, Gonzalo Álvarez (2009), *Plan Director Informático*, MIDES.

[2] Ideasoft Technology & Business Performance
http://www.ideasoft.biz/
Último acceso: Noviembre 2012

[3] W.H. Inmon (2002). *Building the Data Warehouse*. Wiley Computer Publishing, third edition.

[4] Elzbieta Malinowski, Esteban Zimányi (2008). *Advanced Data Warehouse Design*. Springer, ISBN- 978-3-540-74404-7.

[5] Verónika Peralta (2001). Diseño Lógico de Data Warehouses a partir de Esquemas Conceptuales Multidimensionales. Tesis de Maestría. . InCo – Pedeciba, Facultad de Ingeniería de la UdelaR.

[6] Ralph Kimball, Margy Ross (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. Wiley Computer Publishing, second edition.

[7] Daniel Moody, Mark Kortink (2000). *From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design*. DMDW'00, 2000.

[8] InCo (2011). *Material Curso de Diseño y Construcción de Data Warehouse*, Instituto de Computación, Facultad de Ingeniería de la UdelaR.

[9] Erhard Rahm, H. Hai Do (2000). *Data Cleaning: Problems and Current Approaches*. Universidad de Leipzig, Alemania: IEEE Data Engineering Bulletin, Vol. 23(4): 3-13, 2000.

[10] Carlo Batini, Monica Scannapieca (1998). *Data Quality: Concepts, Methodologies and Techniques*, Springer, ISBN-13 978-3-540-33172-8.

[11] Fernando Carpani (2000). *CMDM: Un Modelo Conceptual para la Especificación de Bases Multidimensionales*. Tesis de Maestría. InCo – Pedeciba, Facultad de Ingeniería de la UdelaR.

[12] Planes y programas del MIDES.
http://www.mides.gub.uy/innovaportal/v/14273/3/innova.front/planes_y_programas
Último acceso: Noviembre 2012

[13] Ralph Kimball (2001). *What not to do.*
http://www.kimballgroup.com/2001/10/24/what-not-to-do/
Último acceso: Setiembre 2012

[14] Comisión Interinstitucional Central del Componente Alimentario del Gabinete Social (2012). *Informe Tarjeta Social, Componente Alimentario*. MIDES, MTSS-INDA, MSP, ASSE.

[15] Ralph Kimball (1996). *Monster Dimensions: Design solutions for handling changes in very large dimensions*.
http://www.kimballgroup.com/html/articles_search/articles1996/9605d05.html
Último acceso: Agosto 2012

[16] Ralph Kimball, Margy Ross (2010). *The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence*. Wiley Computer Publishing.

[17] Adriana Marotta (2010). *Material Curso Calidad de Datos*, Instituto de Computación, Facultad de Ingeniería de la UdelaR.

ANEXO 1: Prototipo de aplicación web para la categorización de productos

A continuación se brindan detalles del funcionamiento y la configuración de la aplicación para la categorización del producto.

Configuraciones

Para el desarrollo del prototipo se utilizaron las mismas tecnologías que actualmente se usan en el Ministerio de Desarrollo Social.

- Lenguaje de programación: Java 1.6
- Servidor de aplicaciones: JBoss 5.1.0
- Framework para la presentación: RichFaces 3.3.2
- Base de datos: SQL Server 2005
- Búsqueda aproximada: Fuzzy Lookup de SQL Server Integration Services

Para configurar el servidor de aplicaciones basta con agregar la librería *sqljdbc4.jar* (driver de sql) para conectarse a SQL Server. Luego debe existir un data-source con el nombre *conexión-categoria* que indique la base de datos del DW y las credenciales de un usuario con acceso a la misma.

Ejemplo:

```
<local-tx-datasource>
  <jndi-name>conexion_categoria</jndi-name>
  <connection-url>jdbc:sqlserver://servidor;databaseName=nombreBaseDatos</connection-url>
   <driver-class>com.microsoft.sqlserver.jdbc.SQLServerDriver</driver-class>
   <user-name>usuario</user-name>
   <password>pass</password>
   <min-pool-size>1</min-pool-size>
   <max-pool-size>3</max-pool-size>
   <idle-timeout-minutes>10</idle-timeout-minutes>
</local-tx-datasource>
```

Diseño

En la base de datos del DW se crean dos tablas, *SugerenciaProducto* y *SugerenciaDescartes*, además de las ya existentes *DimProducto* y *DimCategoria*. En la tabla *SugerenciaProducto*

se guardan las sugerencias generadas para una determinada categoría, luego esta tabla es procesada, utilizando la aplicación, por el usuario para determinar si estas sugerencias se van a persistir en la tabla *DimProducto,* agregando productos a una categoría, o si se van a guardar en la tabla *SugerenciaDescartes*. Esta tabla contiene los productos que fueron descartados para esa categoría. Esto se realiza para no tenerlo en cuenta para las próximas sugerencias de esa categoría.

Para generar las sugerencias se construyó un paquete de SSIS utilizando la transformación *Fuzzy LookUp*. La misma sirve para hacer búsquedas por aproximación utilizando algoritmos de Compute Distance. Cuando se desean generar nuevas sugerencias se ejecuta la transformación que busca similitudes entre las descripciones de todos los productos aún sin categorizar y la de los productos ya existentes en la categoría. En caso de que la similitud supere un umbral configurable, estos productos son insertados en la tabla *SugerenciaProducto* para ser ofrecidos como nuevas sugerencias en la aplicación web.

Debido a que la comunicación desde una aplicación web java hacia un paquete de SSIS no siempre resulta fácil de lograr, optamos por crear procedimientos almacenados de SQL Server para que funcionen de intermediarios entre las dos aplicaciones. Estos son *GenerarSugerencias* y *ProcesarSugerencias.*

GenerarSugerencias es el encargado de ejecutar el paquete SSIS, recibe un parámetro indicando el grado de similitud que tienen que tener como mínimo las descripciones de los productos para ser incluidos en las sugerencias.

ProcesarSugerencias toma las sugerencias que el usuario marcó como correctas y actualiza la tabla de productos para incluirlos en la categoría con la que se está trabajando. A su vez agrega a la tabla de descartes las sugerencias rechazadas por el usuario para que no se vuelvan a mostrar.

Por último la aplicación web consta de una interfaz en donde el usuario puede marcar las sugerencias que son correctas y las que no. Estas preferencias son persistidas en la base de datos para posteriormente invocar a los procedimientos almacenados que procesan y generan nuevas sugerencias.

Funcionamiento

A continuación se detallan los pasos a seguir en un ejemplo de categorización.

La categoría de ejemplo que vamos a crear es Vestimenta.

El primer paso es categorizar manualmente los primeros productos que van a generar las sugerencias iniciales. La aplicación cuenta con una búsqueda que permite explorar todos los productos. En esta etapa lo ideal es buscar la mayor cantidad de palabras claves que

sean representativas de la categoría, de esta manera la categorización puede resultar más efectiva.

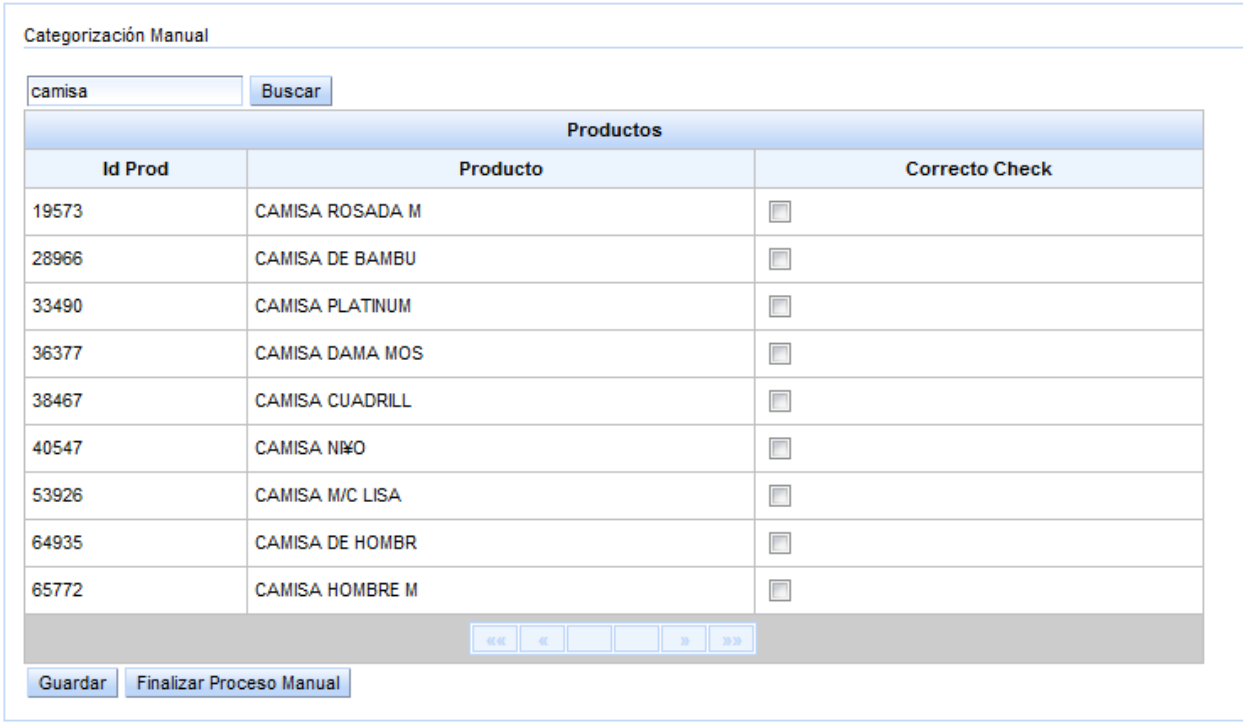
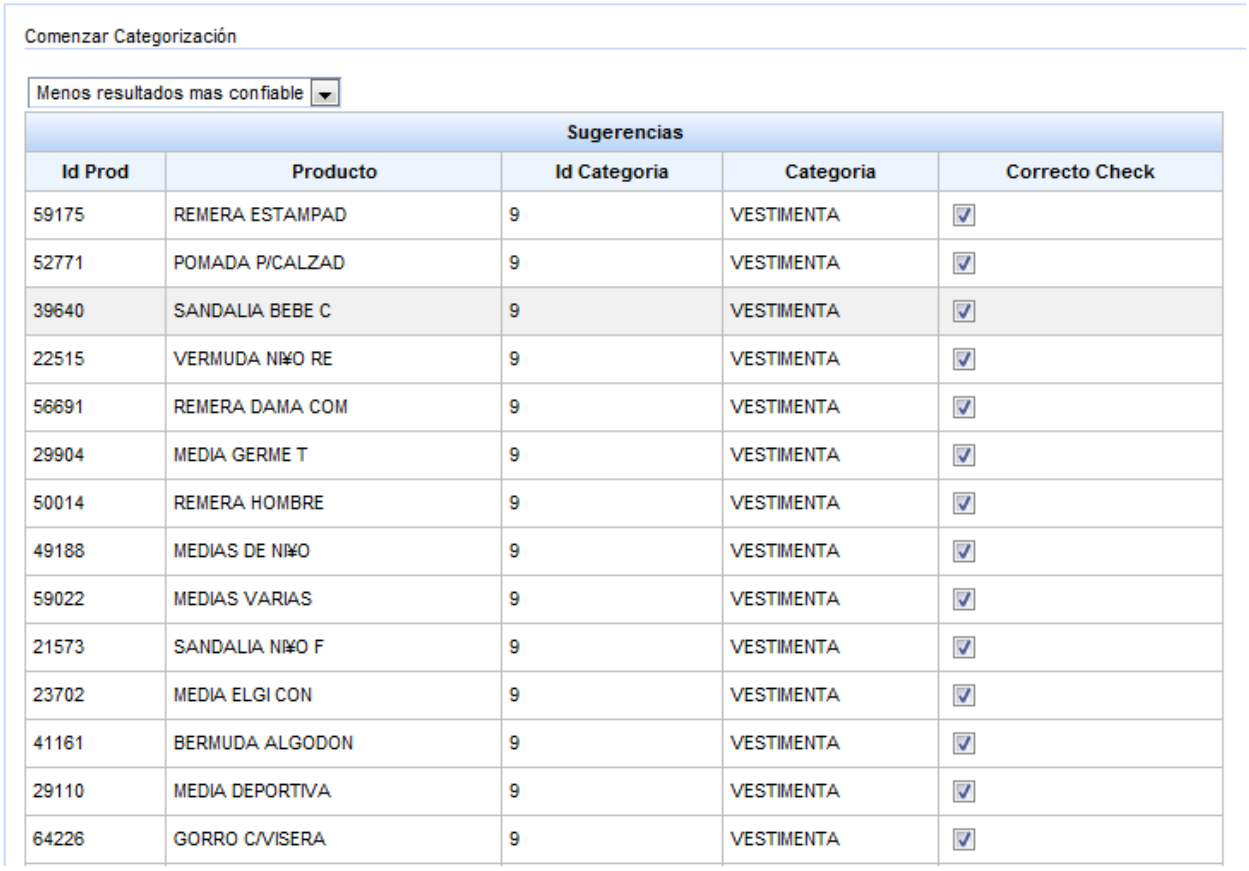


Figura 1: Búsqueda inicial

Luego de categorizar varios productos claves se pasa a la categorización por sugerencias. Se elije primero el grado de similitud con el que se quiere buscar de entre 3 opciones, “Más resultados, menos confiable”, “Medio” y “Menos resultados, más confiable”. La aplicación muestra una lista con sugerencias y el usuario debe indicar cuáles son correctas y cuáles no.



Sugerencias

Id Prod	Producto	Id Categoria	Categoria	Correcto Check
59175	REMERA ESTAMPAD	9	VESTIMENTA	☑
52771	POMADA P/CALZAD	9	VESTIMENTA	☑
39640	SANDALIA BEBE C	9	VESTIMENTA	☑
22515	VERMUDA NI¥O RE	9	VESTIMENTA	☑
56691	REMERA DAMA COM	9	VESTIMENTA	☑
29904	MEDIA GERME T	9	VESTIMENTA	☑
50014	REMERA HOMBRE	9	VESTIMENTA	☑
49188	MEDIAS DE NI¥O	9	VESTIMENTA	☑
59022	MEDIAS VARIAS	9	VESTIMENTA	☑
21573	SANDALIA NI¥O F	9	VESTIMENTA	☑
23702	MEDIA ELGI CON	9	VESTIMENTA	☑
41161	BERMUDA ALGODON	9	VESTIMENTA	☑
29110	MEDIA DEPORTIVA	9	VESTIMENTA	☑
64226	GORRO C/VISERA	9	VESTIMENTA	☑

Figura 2: Sugerencias

En esta lista vemos parte de las sugerencias para la categoría Vestimenta. El usuario deja marcadas con un check las que cree correctas y procesa las sugerencias. Esto hace que se actualicen los productos seleccionados como nuevos productos de la categoría y se vuelve a realizar el proceso de generar sugerencias tomando en cuenta ahora, los productos recientemente agregados. Este proceso se repite hasta que el usuario se quede sin sugerencias y decida dar por terminada la categorización o volver a la búsqueda inicial para agregar manualmente nuevos productos que generen nuevas sugerencias.

ANEXO 2: Procesos ETL

Este anexo contiene los flujos de carga de dimensiones y tablas de hechos que no están contenidos en el informe.

Dimensiones de Uruguay Integra

Localidad

Localidad es una dimensión compartida por ambos programas. La Figura 1 describe el flujo de carga.

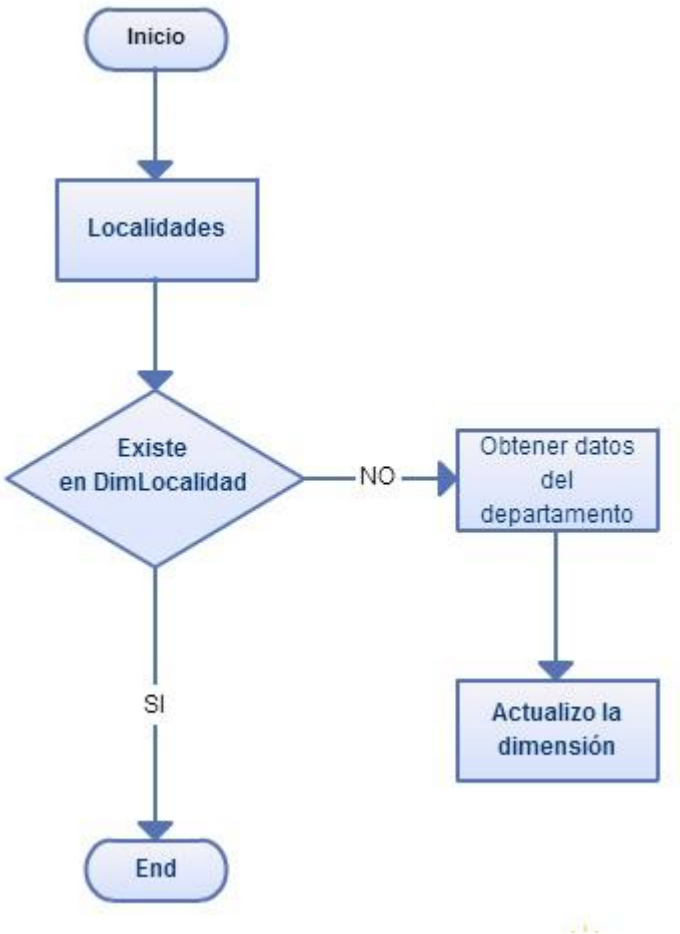


Figura 1: Flujo de carga: dimensión *Localidad*.

Grupo

La Figura 2 muestra el flujo de carga para la dimensión *Grupo*.

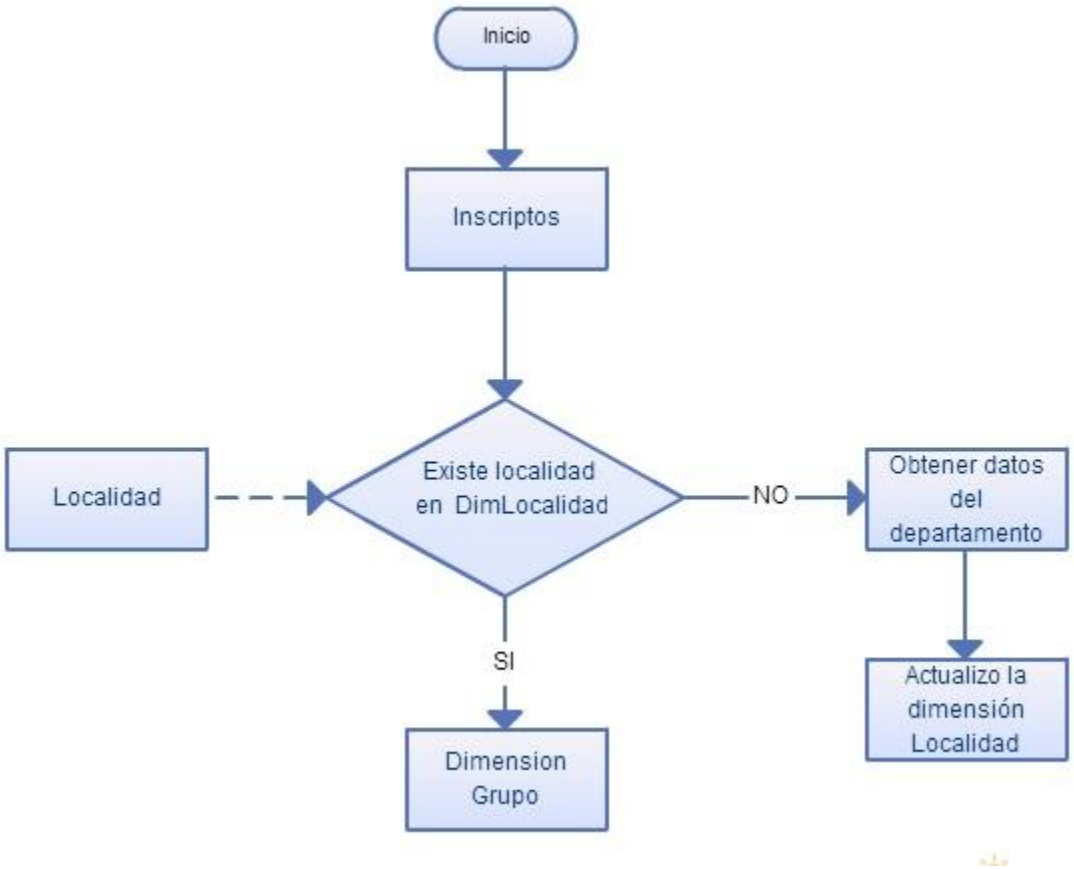


Figura 2: Flujo de carga: dimensión *Grupo*.

Motivo

La Figura 3 muestra el flujo de carga para la dimensión *Motivo*.

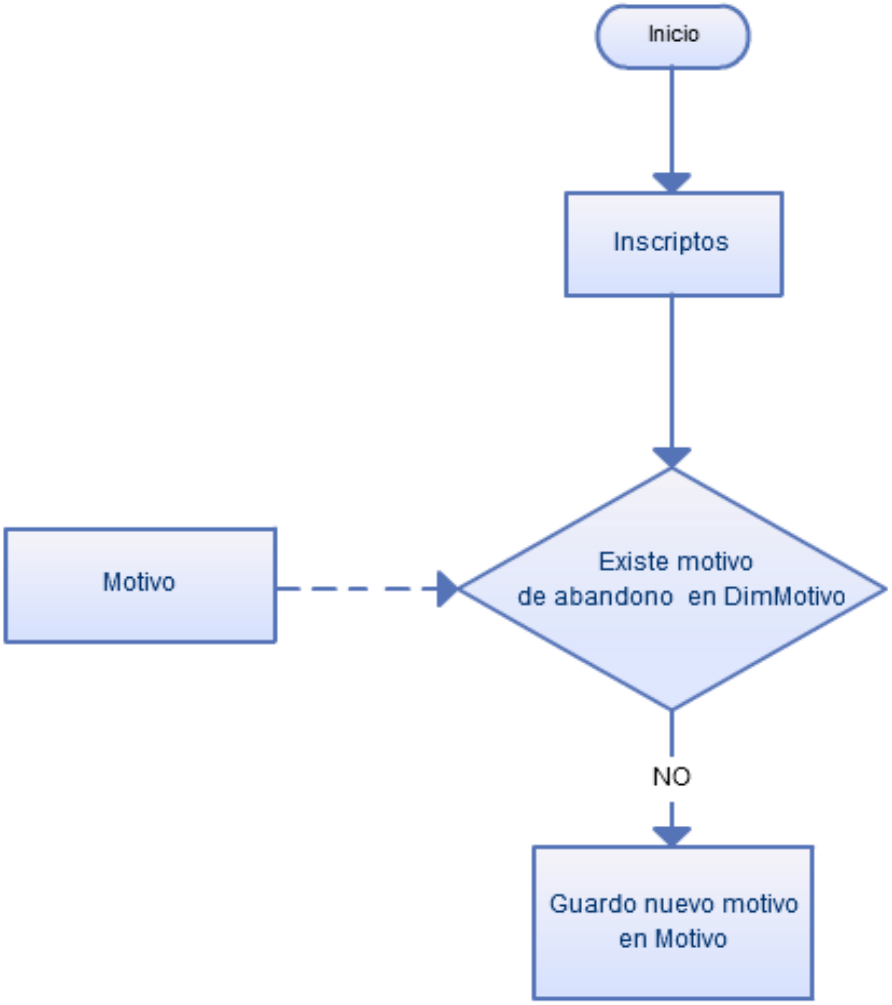


Figura 3: Flujo de carga: dimensión *Motivo*.

Tablas de hechos de Uruguay Integra

Recursos

La Figura 4 muestra el flujo de carga para la tabla de hechos *Fact Recursos*.

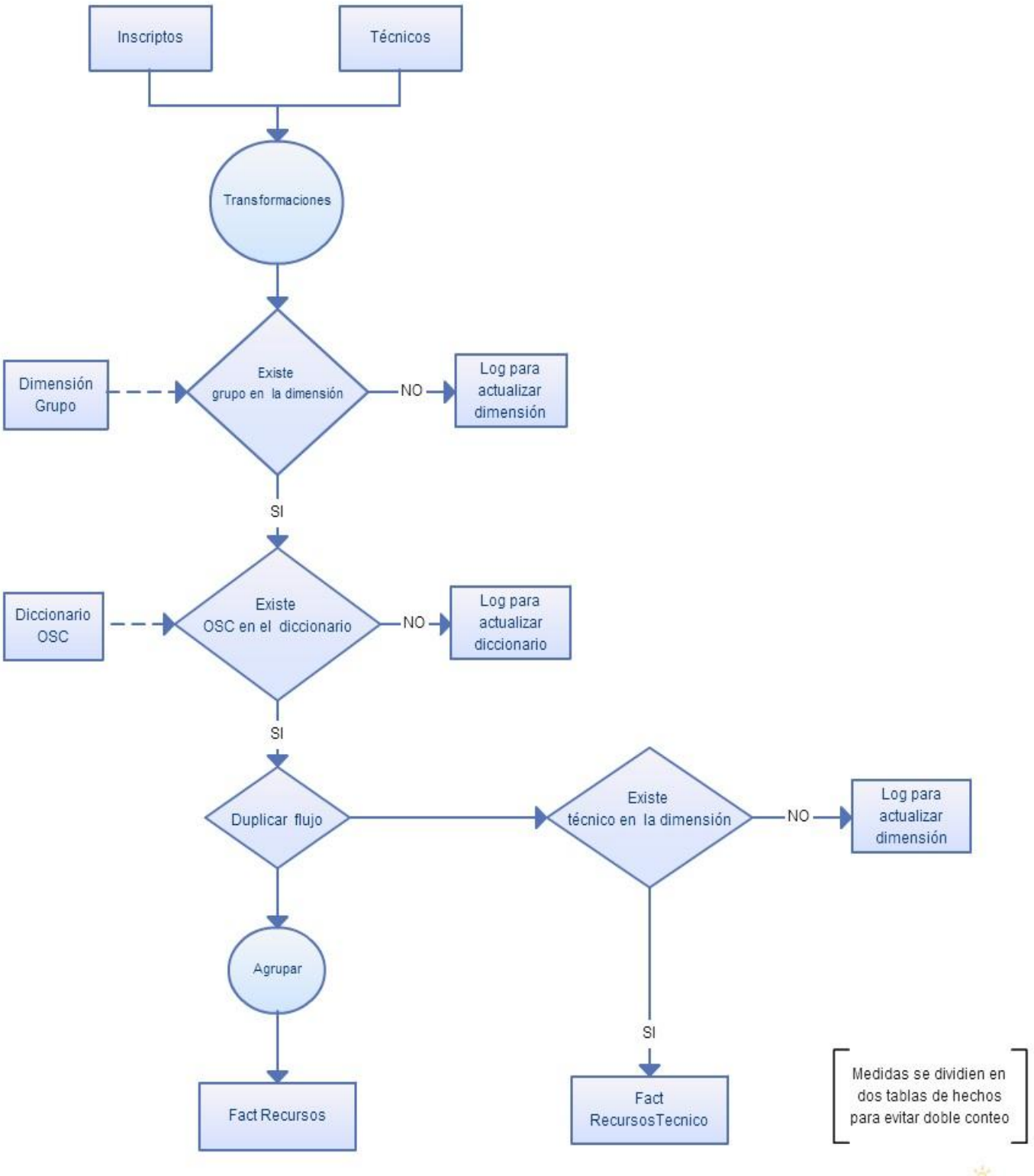


Figura 4: Flujo de carga: Tabla de hechos *Fact Recursos.*

Producto

La Figura 5 muestra el flujo de carga para la tabla de hechos *Fact Producto*.

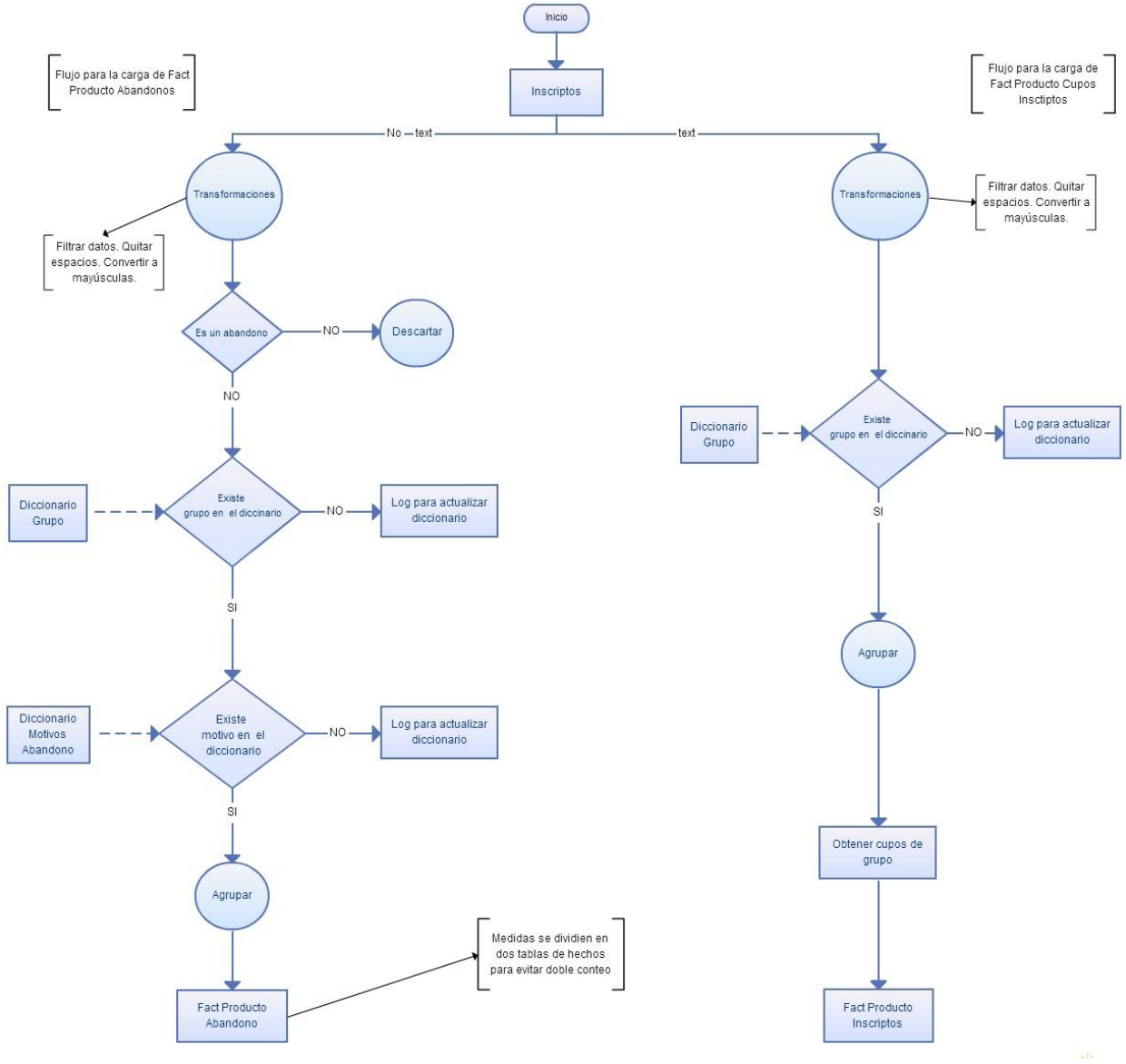


Figura 5: Flujo de carga: Tabla de hechos *Fact Producto.*

Dimensiones de Tarjeta Uruguay Social

Comercio

La Figura 6 muestra el flujo de carga para la dimensión *Comercio*.

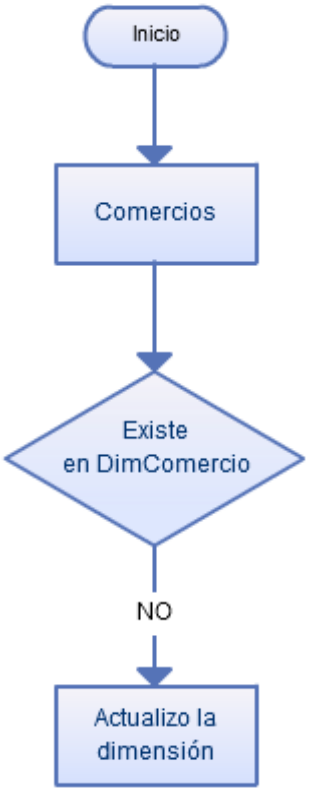


Figura 6: Flujo de carga: dimensión *Comercio*.

ANEXO 3: Base de datos fuente de Tarjeta Uruguay Social

La Figura 1 muestra el diagrama completo de la base de datos fuente de Tarjeta Uruguay Social.

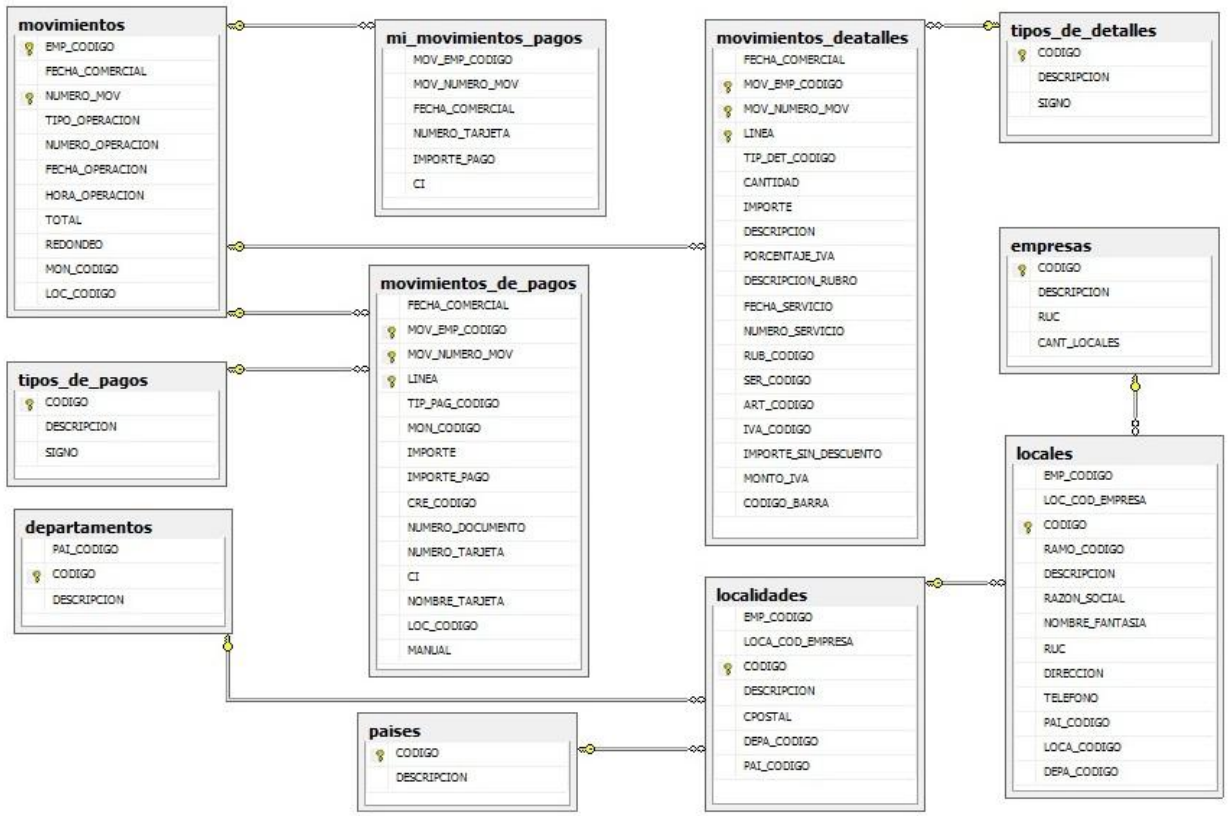


Figura 1: Diagrama de la base de datos fuente de Tarjeta Uruguay Social.