3C Tecnología *Vol.3 – Nº 1 | Febrero – mayo 2014, 8 - 20*
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 4143

Recepción: 16 de septiembre de 2013
Aceptación: 20 de diciembre de 2013
Publicación: 25 de febrero de 2014

MERCADOS DE DATOS PARA EL ANÁLISIS ESTADÍSTICO DE LA INFORMACIÓN.

DATAMARS FOR STATISTICAL ANALYSIS OF THE INFORMATION

Yisel de los Ángeles González Pompa [1]

María Teresa Rosales González [2]

1. Universidad de Granma Facultad Regional Granma Departamento Web y Multimedia, Cuba. E-mail: yisel.pompa@grm.uci.cu
2. Universidad de las Ciencias Informáticas, Cuba.

3C Tecnología Vol.3 – Nº 1 | Febrero – mayo 2014, 8 - 20
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 4143

RESUMEN

La presente investigación se enmarca en el tema de los almacenes de datos, los mercados de datos, y su utilización para los análisis estadísticos. Su desarrollo facilita la toma de decisiones en el área de Cooperación internacional de la Universidad de las Ciencias Informáticas identificando, proyectando y prediciendo tendencias futuras a partir de datos acumulados. Se obtuvo como resultado un mercado de datos poblado y funcional, con una capa de inteligencia de negocio que facilita el estudio de la información.

ABSTRACT

This research is part of the topic of data warehouses, datamarts, and their use for statistical analysis of the information. Its development facilitates decision-making in the area of international cooperation of the University of Information Sciences identifying, planning and predicting future trends based on accumulated data. The result was the development process of a datamart populated data and functional with a business intelligence layer, providing specialists to better study area information.

PALABRAS CLAVE

Almacenes de datos, Cooperación internacional, inteligencia de negocio, mercado de datos, toma de decisiones.

KEY WORDS

Warehouses, International cooperation, business intelligence, decision-making, datamart, decision-making.

INTRODUCCIÓN

El avance alcanzado por las Tecnologías de la Informática y las Comunicaciones ha influido en el desarrollo del ser humano. Continuamente el desarrollo de software amplía su alcance y se gestiona con mayor profundidad. Siendo hoy en día un proceso de simple y rápida ejecución para los usuarios de sistemas informáticos procesar operaciones muy complejas, donde se manejen enormes volúmenes de información. Relacionado a estos sistemas y dentro de su entorno surgen las bases de datos (BD).

Paralelo a ello, la necesidad de extraer información útil a partir de datos históricos se ve reflejada de forma más tangible y valiosa en el mundo de los negocios. Con el objetivo de brindarle soporte a este proceso surgen los sistemas de apoyo a la toma de decisiones, creados para medir y controlar el desarrollo de las variables importantes del negocio. La Universidad de las Ciencias Informáticas (UCI), cuenta con el centro de desarrollo Centro de Tecnologías de Gestión de Datos (DATEC) que se especializa en el desarrollo de soluciones, productos y servicios relacionados con las tecnologías de gestión de datos.

Éste se encuentra colaborando con el proceso de informatización que se está llevando a cabo en la universidad, entre sus proyectos se encuentra el denominado *Sala Situacional UCI* que abarca varias áreas de la universidad, una de ellas es Cooperación internacional, dentro de ésta la información detallada es almacenada en diversos formatos y en períodos variables, convirtiéndose su unificación, manejo e interacción en una tarea compleja. Por lo cual la realización de los reportes del área es de forma manual y con un gran costo de tiempo, lo que trae consigo una demora en el análisis y estudio de la totalidad de la información. Este conjunto de factores trae como consecuencia que los directivos de la UCI no cuenten con las herramientas requeridas para ofrecer datos que respalden las decisiones tomadas, lo que puede incidir de manera negativa en la ejecución de sus procesos regulares y en la prevención de comportamientos futuros que pueden afectar el adecuado funcionamiento de la misma.

3C Tecnología Vol.3 – Nº 1 | Febrero – mayo 2014, 8 - 20
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 4143

CONTENIDO

ALMACENES DE DATOS

Los almacenes de datos (AD) constituyen un escalón superior en la evolución de las BD hacia mayor inteligencia y mejores funcionalidades. Los fundamentos o características de los AD según W. H. Inmon, padre de la disciplina, se basan en que estos sistemas consisten en una colección de datos orientado a materias, integrado, no volátil, que varía con el tiempo y que se diseñan para dar apoyo a los procesos de toma de decisiones [1].

Se puede decir que un AD está orientado a materias porque sus funcionalidades se concentran en un tema en específico o en base a los aspectos que son de interés para la empresa. Además es integrado debido a que se construye mediante la unificación de fuentes de datos múltiples y heterogéneas, eliminando las inconsistencias en el nombrado, estructuras codificadas y medidas de los atributos que puedan surgir entre ellas. Una vez que se encuentra poblado el almacén, los datos no cambian, sus actualizaciones no ocurren en su propio entorno, sino en las fuentes de datos de las que se nutre, siendo ésta una de las razones por las que un AD es no volátil. La variación en el tiempo se ve evidenciada en un AD primeramente porque cada clave contiene una referencia a la fecha explícita o implícitamente. Igualmente el espacio de tiempo en él, es significativamente más largo que el de los sistemas operacionales debido a que la información es mostrada desde una perspectiva histórica [1; 2].

Ventajas

Los AD proporcionan una herramienta para el proceso de toma de decisiones estratégicas y tácticas en cualquier área funcional, teniendo en cuenta información integrada y global del negocio, además aporta la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios [1].

Desventajas

Algunas de las desventajas de los AD son la subestimación de los recursos imprescindibles para la captura, carga y almacenamiento de los datos, el incremento continuo de los requerimientos de información del AD [3], y la subestimación de las funcionalidades que ofrece la correcta utilización del AD [1].

MERCADO DE DATOS

Un MD es la implementación de un almacén con alcance limitado a un área funcional, problema en específico, departamento, tema o grupo de necesidades [1]. Este se puede integrar a un AD de dos formas, dependiendo del enfoque de la arquitectura que se decida implementar:

Enfoque de Inmon

Define previamente el AD y posteriormente delimita sobre él los MD. Esta implementación tiene como principal inconveniente que se requiere una inversión muy costosa en tiempo y recursos [1].

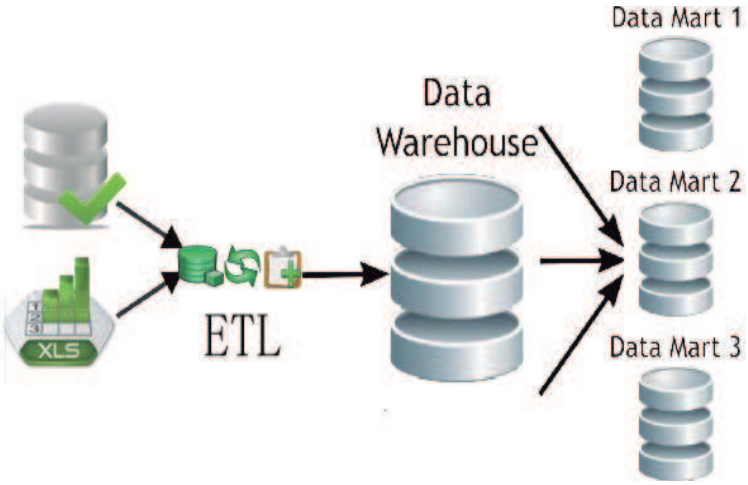

Figura 1 Arquitectura descendente. **Fuente:** Elaboración propia.

Enfoque de Kimball

Plantea crear previamente los MD departamentales y posteriormente integrarlos en un AD para la organización, la fundamental ventaja de esta implementación es que se desarrolla y despliega en un período de tiempo menor que una realizada con el enfoque de Inmon y con menos recursos. El principal problema que se presenta al aplicar esta solución está dado en tener que sincronizar los hechos cuando se realice la consolidación del AD [1].

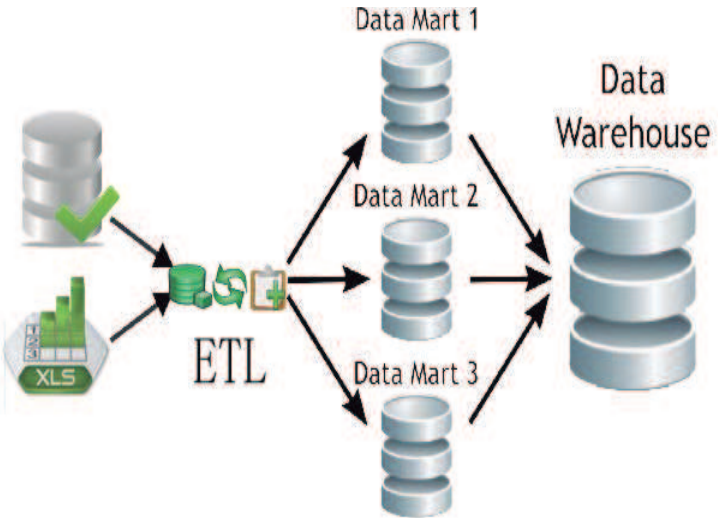

Figura 2 Arquitectura ascendente. **Fuente:** Elaboración propia.

Modelo multidimensional

En un AD se representan los datos usando modelos multidimensionales para facilitar su análisis. Un modelo multidimensional proporciona dos conceptos fundamentales: hecho y dimensión. Constituyendo el cubo el objeto central en el modelado de datos multidimensional.

Un cubo consiste en un conjunto de celdas, de tal manera que cada una está identificada por la combinación de los miembros de las diferentes dimensiones y contiene el valor de la medida analizada para dicha combinación de dimensiones [4].

PROCESOS BÁSICOS DE UN ALMACÉN DE DATOS

Para el desarrollo exitoso de un AD se requiere seguir una serie de procesos en cada una de las etapas del mismo. Estas etapas van desde el análisis de la organización hasta el despliegue del AD y básicamente estos procesos se resumen en dos: la integración de los datos y el BI.

Integración de datos

Aunque la creación del ETL constituye una actividad que pasa desapercibida en muchas ocasiones al usuario final, fácilmente consume el 70 por ciento de los recursos requeridos para la implementación y el mantenimiento de un típico AD [5]. En el proceso ETL se añaden valores significativos a los datos. Esto es mucho más que profundizar para obtener datos de los sistemas fuentes y del AD. Fundamentalmente el ETL permite realizar funcionalidades como eliminar errores y corregir datos faltantes, capturar el flujo de datos transaccionales para su protección, concordar los datos de múltiples fuentes para ser usadas conjuntamente, estructurar los datos para ser usables por las herramientas del usuario final [6].

Inteligencia de negocio

En la actualidad la información se ha convertido en fundamento para conseguir superioridad competitiva en los negocios. Las funcionalidades de las herramientas de BI están encaminadas a generar reportes, gráficos en diferentes perspectivas, vistas de análisis y todo lo que se necesite para que el usuario final pueda analizar y razonar los datos con mayor facilidad, con el objetivo de llegar a conclusiones correctas con mayor prontitud.

El BI constituye la habilidad corporativa para tomar decisiones. Esto se logra mediante el uso de metodologías, herramientas y tecnologías que permiten reunir, depurar, transformar datos, y aplicar en ellos técnicas analíticas de extracción de información. Las herramientas BI permiten mayor rapidez en la obtención y lectura de la información, y en consecuencia para la toma de decisiones, visibilidad inmediata y automática de los nuevos datos incorporados [7].

METODOLOGÍA DE DESARROLLO DE ALMACENES DE DATOS

Una metodología es un conjunto de pasos o acciones definidas para guiar un proceso determinado. En el caso del desarrollo del software, su definición no se aleja de este marco,

a esto se le añade la definición de las etapas, los roles y las tareas para la generación de los artefactos finales de productos informáticos [8]. Con el transcurso de la madurez de los AD, se han convertido en referente en cuanto a sistemas para el apoyo en la toma de decisiones partiendo de información histórica. Las más reconocidas en el mundo son la metodología de Kimball y la metodología de Inmon ambas abundantemente documentadas y definidas.

Para las soluciones de esta rama DATEC propone una la Metodología para el desarrollo de almacenes de datos en DATEC que abarca todas las fases del desarrollo de un AD, esta toma como fundamento la metodología de Kimball en los siguientes teniendo en cuenta que ésta propone los conceptos de hechos y dimensiones, lo cual apoya la toma de decisiones y también en el proceso de desarrollo. Plantea que la construcción del AD sea a medida que se construyan las BD departamentales, lo que da la impresión de organización en la empresa, entidad u organización. Abundante documentación sobre la misma, posibilitando así la aclaración de las dudas que vayan surgiendo relacionadas a ese tema. Amplio reconocimiento por los especialistas que se desempeñan en esta disciplina, pues constituye una metodología madura y que tiene bien definidas las etapas, actividades, artefactos y roles [6].

RESULTADOS Y DISCUSIÓN

En la presente sección se muestran los resultados de la aplicación de la metodología y los elementos que se tuvieron en cuenta en cada una de las fases de desarrollo. Así como los resultados obtenidos al probar el mercado de datos una vez poblado y funcional. Además se evidencian las experiencias adquiridas y las ventajas que puede ofrecer la utilización de la aplicación.

ANÁLISIS, DISEÑO E IMPLEMENTACIÓN DEL MD COOPERACIÓN INTERNACIONAL

En el proceso de análisis se llevó a cabo la traducción de las necesidades de los usuarios, es decir lo que precisan que el sistema les facilite, en requisitos. De la calidad con la cual se realiza la extracción de estas necesidades depende en gran medida el éxito o fracaso del software.

El análisis comprende una serie de tareas entre las que se encuentra la definición de las reglas de negocio (RN) que describen el nivel adecuado de detalle hasta el que se va a llegar en la implementación. Para la presente investigación se declaró que una vez que los datos estén cargados en el almacén, no pueden existir campos nulos, además el código de los atributos en cada una de las dimensiones no puede tomar valores repetidos y no pueden existir campos con valores negativos.

Otra de las actividades es la definición de los 15 requisitos de información (RI), estos consisten en el conjunto de funcionalidades que se deben implementar para analizar los datos en cuanto a parámetros previamente definidos por el cliente.

Por último se definen los requisitos funcionales (RF) y los requisitos no funcionales (RNF), según la Ayuda del Proceso Unificado de Software (RUP) un RF es una capacidad de software necesaria para que el usuario solucione un problema al alcanzar un objetivo [9].

ARQUITECTURA DEL MERCADO DE DATOS

La arquitectura de la solución se define a partir de los RNF obtenidos en el levantamiento de requisitos. Esta comprende aspectos como, la comunicación entre los subsistemas y la tecnología a utilizar, entre otros aspectos de gran importancia [10]. Para el caso del MD Cooperación internacional la arquitectura definida consta de tres subsistemas: integración de datos, almacenamiento y visualización.

El subsistema de integración de datos extrae los datos desde las distintas fuentes, en este caso excel y BD, comunicándose con la última mediante el protocolo TCP/IP. Luego se encarga de realizar la limpieza, transformación e integración de los mismos mediante las herramientas Kettle, del PDI 4.2.1 y DataCleaner 1.5.3, dejándolos listos para la carga en el MD.

El subsistema de almacenamiento toma como elementos de partida los datos manipulados por el subsistema de integración utilizando el protocolo de comunicación TCP/IP. En él se

encuentra el modelo físico propuesto soportado en el SGBD PostgreSQL en su versión 9.1 y mediante la utilización de la herramienta PgAdminIII 1.14.0 se garantizará la interacción de los clientes de administración con éste.

Los datos que están en el subsistema de almacenamiento son consultados por el subsistema de visualización mediante protocolo TCP/IP con la herramienta Pentaho BI server 3.10. A este subsistema es que acceden los distintos clientes para obtener los reportes deseados. Para un mejor entendimiento de lo antes expuesto ver figura 3.

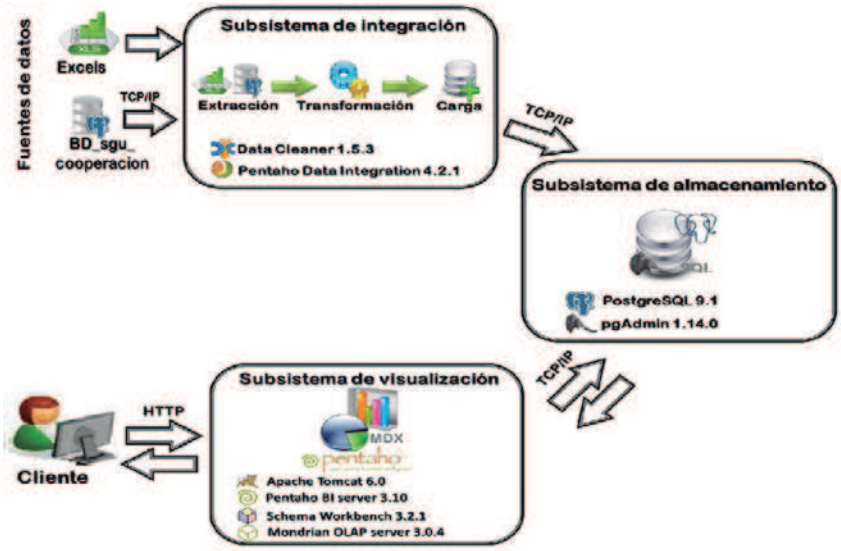

Figura 3 Arquitectura del MD Cooperación internacional

DISEÑO DEL SUBSISTEMA DE INTEGRACIÓN DE DATOS

Para garantizar que la implementación del subsistema de integración de datos se realice de forma exitosa primero se debe concebir un diseño que lo soporte, teniendo en cuenta la mayor cantidad de aspectos y con el mayor nivel de detalle posible, este comprende aspectos como la definición de la arquitectura de integración para definir cómo se debe realizar el proceso de ETL, dónde se incluyen aspectos como los tipos de archivos fuente y su interacción con las herramientas dentro del entorno de trabajo y por último la gestión de metadatos para controlar la ejecución de los procesos de integración de datos.

Otro aspecto importante a tener en cuenta en este proceso el perfilado de datos, éste constituye una de las primeras tareas en el proceso de calidad de datos, consiste en ejecutar un primer análisis sobre los datos fuentes, usualmente sobre las tablas, con el objetivo de comenzar a conocer su estructura, formato y nivel de calidad. Al realizar el perfilado de los datos a las fuentes del MD Cooperación internacional se identificó que los tipos de datos de las fuentes son en su mayoría varchar, float e integer, pero además existe la presencia de fechas con formato timestamp. Además se encontraron 612 valores nulos, no se hallaron valores vacíos ni negativos.

3C Tecnología *Vol.3 – Nº 1 | Febrero – mayo 2014, 8 - 20*
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 4143

DISEÑO DEL SUBSISTEMA DE VISUALIZACIÓN

El subsistema de visualización constituye un elemento fundamental para el MD, debido a que el usuario final interactúa con él. Por ello realizar un correcto y detallado diseño del mismo brinda muchas facilidades en la implementación. Es por ello que definir una arquitectura de información que responda a las particularidades del área brinda la posibilidad de organizar los contenidos para que el acceso a ellos se pueda realizar de forma más simple y directa.

IMPLEMENTACIÓN DEL MERCADO DE DATOS COOPERACIÓN INTERNACIONAL

Como parte de la implementación del subsistema de almacenamiento se desarrolló el modelo físico, éste constituye una colección integrada de entidades que describe las estructuras de los datos. Dicho modelo se genera a partir del modelo lógico dimensional; conteniendo las relaciones entre las tablas de hechos y dimensiones que conforman el MD.

En el subsistema de integración de datos incluye aspectos como la gestión del cambio en las dimensiones para los datos que tienden a ser modificados en el transcurso del tiempo. Fueron utilizados los metadatos de proceso con el objetivo de obtener la información de las transformaciones y los trabajos pertinentes a los subprocesos de ETL. Se realizaron 11 transformaciones para la carga de los hechos, en las cuales se utilizaron un conjunto de subsistemas que Kimball propone para la implementación del proceso de ETL y los trabajos.

Por último se implementó el subsistema de visualización, proceso muy importante en el desarrollo de un MD, debido a que define qué información se presenta para los usuarios y cómo se muestra la misma para el análisis. Como parte de la implementación del mismo se desarrollaron un total 11 cubos OLAP, uno por cada hecho. Estos permiten estructurar los datos para que concuerden con el modo que tienen los usuarios de analizarlos, luego se realizaron los reportes teniendo en cuenta los RI.

PRUEBAS REALIZADAS AL MERCADO DE DATOS COOPERACIÓN INTERNACIONAL

Para validar el correcto funcionamiento de un software es importante ir probándolo desde el inicio de su construcción hasta después de aceptado por el cliente. Las actividades de control de la calidad del mismo deben tener un orden de ejecución lógico y funcional, por lo cual el centro DATEC define el uso del modelo V con estos fines.
Como parte de la estrategia de prueba para aplicar los tipos de pruebas mencionados anteriormente al MD Cooperación internacional, se emplearon como herramientas los casos de prueba (CP) para CU y reglas de transformación, las listas de chequeo a los artefactos de ETL y el perfilado de los datos destino para probar la calidad de los mismos. Para el MD Cooperación internacional fueron diseñados cuatro CP. En el proceso de ETL también se aplicaron CP, en este caso para las reglas de transformación con la finalidad de comprobar si luego de ejecutada cada transformación, los datos cargados sean los

esperados, encontrándose un grupo de inconformidades resueltas satisfactoriamente. Luego fueron aplicadas las listas de chequeo por los especialistas de calidad.

PRUEBAS UNITARIAS Y DE INTEGRACIÓN

Al concluir la implementación del sistema se le aplicaron pruebas unitarias a los flujos de integración de datos y a la capa de visualización, arrojando como resultado tres no conformidades en cada caso, para un total de seis, las cuales fueron resueltas satisfactoriamente.

PRUEBAS DE SISTEMA

Las pruebas del sistema al MD Cooperación internacional se aplicaron por parte del grupo de calidad del departamento de AD, encontrándose tres no conformidades de complejidad alta, todas resueltas satisfactoriamente. Quedando el sistema liberado para su despliegue en el área.

PRUEBAS DE ACEPTACIÓN

En conjunto con el cliente se realizaron las pruebas de aceptación de la solución, las cuales arrojaron resultados satisfactorios, quedando comprobado que el sistema cumple con sus necesidades y que están satisfechos con el producto elaborado.

CONCLUSIONES

Con la realización de la investigación se logró cumplir con los objetivos planteados, desarrollándose un mercado de datos poblado y funcional que ofrece soporte a la toma de decisiones en el área de Cooperación internacional de la Universidad de las Ciencias Informáticas. Por lo que se concluye que:

1. El estudio de los fundamentos teóricos de la investigación, permitió seleccionar una metodología para organizar de manera estructurada el proceso de desarrollo de software.
2. El análisis y diseño del mercado de datos, posibilitó un mejor entendimiento del proceso de negocio de la dirección de Cooperación internacional permitiendo la implementación de una solución que responde a las necesidades del cliente.
3. La implementación del mercado de datos contribuye al proceso de apoyo a la toma de decisiones mediante el desarrollo de los diferentes subsistemas, garantizado la correcta organización, carga y visualización de los datos.
4. La validación de la solución mediante la aplicación de los casos de pruebas, el perfilado de los datos y las listas de chequeo permitió obtener un mercado funcional que cuenta con una correcta calidad de los datos y que cumple con los requisitos especificados por el cliente.

REFERENCIAS

1. Bernabeu, Ricardo Dario. *HEFESTO: Metodología propia para la Construcción de un Data Warehouse.* Córdoba, Argentina : s.n., 2007.
2. Giménez, Matilde Celma. *Almacenes de Datos. Departamento de Sistemas Informáticos y Computación*, Universidad Politécnica de Valencia. España : s.n. 2008.
3. Zorrilla, Marta. *Data warehouse y OLAP*. Cantabria, España. 2007 – 2008.
4. Dapena Bosquet, Isabel. Muñoz. *Sistemas de Información Orientados a la Toma de Decisiones:el enfoque multidimensional.* Madrid, España: s.n., 2005.
5. Kimball, Ralph and Caserta, Joe. *The Data Warehouse ETL Toolkit*. Canadá : Wiley Publishing, 2004.
6. Kimball, Ralph and Ross, Margy. *The Datawarehouse Toolkit.* Second Edition. s.l.: John Wiley & Sons, Inc, 2002. Vol. *The Complete Guide to Dimensional Modeling*.
7. Inmon, W. *Building the Data Warehouse*. 2nd edition. John Wiley & Sons. 1996.
8. González Hernández, Ing.Yanisbel. *PROPUESTA DE METODOLOGIA PARA EL DESARROLLO DE ALMACENES DE DATOS EN DATEC*. 2012.
9. Corporation, IBM. *Ayuda de Rational Unified Process*. [Online] 7.0.1, 1987-2006.
10. Pressman, Roger S. *Ingería del software. Un enfoque práctico.* s.l. : McGrawHill. Vol. 6ta edición. ISBN-970-10-5473-3.
11. González Pompa, Yisel de los Angeles y Rosales González, Maria Teresa. 2012. *Mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas.*: s.n., 2012.