Revista Tekhnê
2011, Vol. 8, 55–58

ISSN 1692-8407

Sistema bodega de datos para la administración de los indicadores del Sistema de Universidades Estatales (SUE) soportado por un sistema distribuido, para la Universidad Distrital Francisco José de Caldas

Data warehouse system for managing the indicators of State University System (SUE) supported by a distributed system for the District University Francisco José de Caldas

Jeisson A. Hernández
Universidad Distrital Francisco José de Caldas
jeissonhernandez@gmail.com

Andrés F. Mora
Avanxo
fmora@avanxo.com

El presente documento describe el proceso de análisis, diseño e implementación de un Sistema Bodega de Datos basado en Inteligencia de Negocios y soportado por un Sistema Distribuido para la Universidad Distrital Francisco José de Caldas. El sistema fue desarrollado teniendo en cuenta la Metodología *RoadMap* de inteligencia de negocios y la metodología *Top-Down* de sistemas distribuidos; Además de esto, se implementaron herramientas de procesamiento analítico en línea OLAP (*On Line Analytical Processing*), procesos ETL (Extracción, Transformación y Carga) y reportes para realizar el análisis de los indicadores.

Palabras clave: bodega de datos, OLAP, RoadMap, sistemas distribuidos, SUE

This document describes the process of analysis, design and develops of a data warehouse system based on Business Intelligence and supported by a Distributed System for Universidad Distrital Francisco José de Caldas. The system was developed taking into account the RoadMap methodology of business intelligence and Top-Down methodology for distributed systems; addition, tools were implemented online analytical processing OLAP, ETL (Extract, Transform and Load) and reports for analysis of indicators.

Keywords: data warehouse, distributed systems, OLAP, RoadMap, SUE

Introducción

A nivel normativo en Colombia se ha establecido un sistema para evaluar los procesos de gestión de las Universidades Estatales, lo que dio origen al Sistema de Universidades Estatales u Oficiales (SUE). El sistema fue creado por la Ley 30 de 1992 (artículo 81), para elaborar planes periódicos de desarrollo institucional, considerando las estrategias de planeación regional y nacional. Una de sus posibilidades de acción establece un sistema común de indicadores de gestión (Sistema Universitario Estatal (SUE), 2001). La Universidad Distrital debe reportar año tras año los indicadores establecidos por el SUE, lo cual en la actualidad es una tarea poco eficiente, debido a que la Universidad no tiene su información centralizada, ni sistematizada.

Como solución, en el presente documento se plasma el proceso de diseño e implementación del *Sistema Bodega de Datos para la administración de los indicadores del Sistema de Universidades Estatales (SUE), soportado por un sistema distribuido, para la Universidad Distrital Francisco José de Caldas*. Como productos finales presentará reportes y vistas de análisis que responderán a los requerimientos realizados por algunos de los indicadores planteados por el SUE. El Sistema Bodega de datos no dio respuesta a la totalidad de los indicadores del SUE debido a que parte de la información solicitada no existe en la Universidad. Por otro lado, la in-

Fecha recepción del manuscrito: Agosto 30, 2011
Fecha aceptación del manuscrito: Octubre 30, 2011

Jeisson A. Hernández, Facultad Tecnológica, Universidad Distrital Francisco José de Caldas; Andrés F. Mora, Avanxo.

Esta investigación fue financiada por: Universidad Distrital Francisco José de Caldas.

Correspondencia en relación al artículo debe ser enviada a: Andrés F. Mora. Email: andresmora20@gmail.com

formación existente está en un nivel bajo de sistematización y solo algunas dependencias brindaron acceso limitado a su información (por políticas de seguridad de las mismas dependencias) por medio físico o magnético y vistas a las bases de datos de sus sistemas de información.

A nivel técnico, se tratan temas como la inteligencia de negocios (Cano, 2007), la arquitectura de inteligencia de negocios, bodegas de datos, OLAP, ETL. Dichos sistemas se caracterizan por estar distribuidos con sus respectivas ventajas, desventajas y propiedades. Para finalizar, en la parte conceptual de indicadores, se mostrará su definición, clasificación y sistemas de indicadores enfatizando en el Sistema Universitario Estatal SUE.

El Sistema Bodega de Datos fue construido, para su contenido en Inteligencia de Negocios, sobre la metodología *RoadMap* (la cual consiste en llevar el proyecto a través de las seis fases de un proyecto de ingeniería, cumpliendo con 16 pasos principales) y para la implementación del sistema distribuido la metodología *Top-Down*. El desarrollo de las anteriores metodologías se presentará a lo largo del documento.

Metodología

Teniendo en cuenta que todos los proyectos de ingeniería deben atravesar seis fases entre la creación y la implementación, siendo iterativo el proceso, que a su vez es mejorado por la retroalimentación, el proyecto sistema bodega de datos no se aleja de esta premisa (Kimball y Ross, 2002). Por tal razón, se decidió que para el desarrollo del proyecto se combinaran dos metodologías, dada la complejidad del mismo. La primera hace referencia a la construcción del sistema bodega de datos y se basa en la Metodología *RoadMap*. Para el sistema distribuido, se utilizará la metodología *Top-Down*.

La metodología *RoadMap* consiste en llevar el proyecto de BI (*Business Intelligence*) por diferentes etapas que garanticen la expansión del mismo. Está envuelta en las seis etapas comunes de un proyecto de ingeniería, y con 16 pasos para su desarrollo. Las etapas son mostradas en la Figura 1. La numeración indicada en ésta figura es explicada en la Tabla 1.

La metodología *Top-Down* consiste en fijar unos criterios y especificaciones iniciales del proyecto en un nivel jerárquico superior. Estas especificaciones de nivel superior son sucesivamente transferidas de un modo hereditario a todas las partes del proyecto de los niveles inferiores (Aleixos, Patalano, Contero, Company, y Vila, 2001). La metodología está compuesta por seis fases, cada una con actividades específicas por desarrollar. La integración de las dos metodologías a utilizar para el desarrollo del proyecto es mostrada en la Tabla 1.

Tabla 1
Integración de metodologías

FASE	METODOLOGÍA	ACTIVIDADES
PLANIFICACIÓN	ROADMAP INTELIGENCIA DE NEGOCIOS	1) Evaluación del Caso de Negocio
	TOPDOWN SISTEMA DISTRIBUIDO	- No aplica
JUSTIFICACIÓN	ROADMAP INTELIGENCIA DE NEGOCIOS	2) Infraestructura de la Empresa. 3) Planeación del Proyecto
	TOPDOWN SISTEMA DISTRIBUIDO	- Análisis de las metas y restricciones. - Análisis de los objetivos técnicos
ANÁLISIS	ROADMAP – INTELIGENCIA DE NEGOCIOS	4) Definición de Requerimientos. 5) Análisis de Datos. 6) Prototipo de Aplicación. 7) Análisis del repositorio de Metadato
	TOPDOWN SISTEMA DISTRIBUIDO	- Caracterización de la red existente. - Análisis de la red actual - Análisis para la construcción de Clúster en Linux
DISEÑO	ROADMAP – INTELIGENCIA DE NEGOCIOS	8) Diseño de la Base de Datos 9) Diseño de la ETL. 10) Diseño del Repositorio de Metadatos
	TOPDOWN SISTEMA DISTRIBUIDO	- Diseño de la topología de red - Selección y configuración del sistema operativo - Selección del tipo de Clúster o Middleware - Modelo de nombres y direccionamiento - Diseño de políticas de seguridad
CONSTRUCCIÓN	ROADMAP – INTELIGENCIA DE NEGOCIOS	11) Desarrollo de la ETL, 12) Desarrollo de la Aplicación, 13) Minería de Datos, 14) Desarrollo del Repositorio de Metadatos
	TOPDOWN SISTEMA DISTRIBUIDO	- Instalación y configuración de sistemas operativos - Configuración de servicios y protocolos de red - Instalación y configuración de servidores de aplicaciones - Instalación y configuración de motores de bases de datos - Implementación de políticas de seguridad
DESPLIEGUE	ROADMAP – INTELIGENCIA DE NEGOCIOS	15) Implementación 16) Pruebas
	TOPDOWN SISTEMA DISTRIBUIDO	- Ejecutar pruebas de Stress simulando un ambiente de „n“ usuarios concurrentes para determinar el rendimiento del clúster - Realizar pruebas de tolerancia a fallos y caídas del sistema - Comparar los resultados de rendimiento tomados antes y después de realizado el proyecto y realizar conclusiones técnicas y de desempeño.
GESTIÓN	ROADMAP – INTELIGENCIA DE NEGOCIOS	- No Aplica
	TOPDOWN SISTEMA DISTRIBUIDO	- Monitoreo del diseño de red - Documentación del diseño de red

Resultados

El principal resultado fue la implementación del Sistema Bodega de Datos, en el cual a nivel de inteligencia de negocios tenemos los listados:

- Cuatro fuentes de datos distintas analizadas y extraídas (Sistema de Información Cóndor, Sistema de Información SICIUD, Base de datos Oficina de Docencia, Base de datos Oficina Autoevaluación y Acreditación de Calidad).
- Cuatro *Datamarts* implementados (*Datamart* de Estudiantes, Docencia, Investigaciones y Proyectos Curriculares).

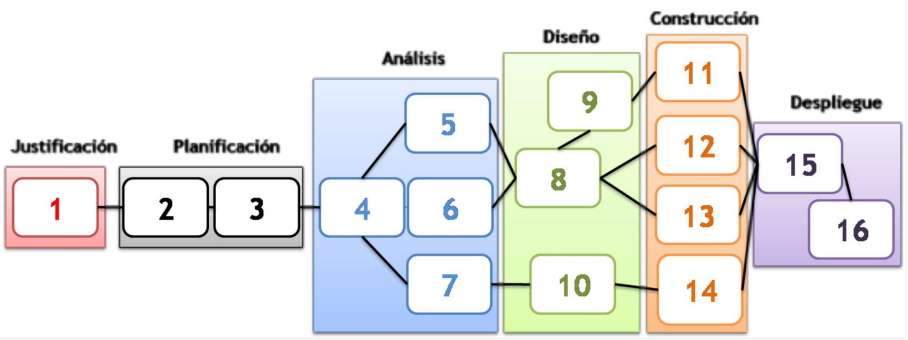

Figura 1. Business Intelligence RoadMap (Moss y Atre, 2003).

- Seis Cubos OLAP diseñados e implementados.
- 29 ETL diseñadas e implementadas.
- Seis *Jobs* de Cargas.
- Seis Vistas de análisis desplegadas y en funcionamiento.
- 37 Reportes diseñados e implementados sobre el Sistema Bodega de Datos SUE discriminados de la siguiente manera: Proyectos Curriculares tres, Grupos y Semilleros de Investigación ocho, Proyectos de Investigación cinco, Estudiantes siete, Docencia nueve, Producción de los docentes cinco.

Otro resultado importante que se debe mencionar es el porcentaje de cubrimiento de los indicadores SUE, reportado a través de la Bodega de Datos. De los 38 indicadores planteados por el SUE, la bodega de datos da respuesta a 24, lo cual significa un cubrimiento del 60,5 % el cual es muy alto teniendo en cuenta las restricciones de acceso a la información o la inexistencia de la información para dar respuesta a algunos de los indicadores.

Por otro lado, en cuanto al sistema distribuido y de acuerdo a la implementación y pruebas de *Stress* realizadas para evaluar el rendimiento, se realizó la comparación entre el sistema bodega de datos soportado por un sistema distribuido y por un sistema centralizado. Como resultado de la comparación a nivel de rendimiento, el Sistema Distribuido, comparado con el Sistema Centralizado a nivel de uso de CPU, tiene un mejor rendimiento, cercano al 50 % en la reducción de uso de CPU. A nivel de uso de memoria RAM, el Sistema Distribuido presenta una mejora de rendimiento visible en una ganancia superior a 250 MB de uso de memoria en su nodo con consumo más alto, comparado con el consumo de memoria del Sistema Centralizado.

En cuanto a la tasa de recepción, el Sistema Centralizado genera cerca de un 26 % más de tráfico sobre la red que el Sistema Distribuido, debido a que gracias a su balanceador de carga asigna las peticiones de los usuarios a los nodos que estén disponibles. Es por esto que en los nodos la tasa de recepción disminuye significativamente debido a las funciones del balanceador.

En lo referente a la tasa de transmisión, el comportamiento del Sistema Distribuido, como del Sistema Centralizado, no difieren significativamente. Cabe destacar que el Sistema Distribuido, según las pruebas de *Stress* realizadas, genera un tráfico mayor a nivel de transmisión debido a que los nodos están continuamente enviando mensajes del estado de los mismos al balanceador.

Dado que una las características de los sistemas distribuidos es la compartición de recursos (Coulouris, Dollimore, Kindberg, y Blair, 2011), el Sistema Bodega de Datos SUE aplica esta característica en la compartición de almacenamiento en disco duro para los *datamarts*.

Para determinar la vida útil del Sistema Bodega de Datos SUE a nivel de almacenamiento en disco de los *datamarts*, se realizó una proyección que definió en cuántos años se excedería la capacidad de almacenamiento de un disco duro de 150 GB, si todos los *datamarts* se almacenaran en un solo servidor y si cada *datamart* tuviera su propio servidor. Para esto se calculó el tamaño en GB por año de cada *datamart*, dando cuenta de una tolerancia del 10 % sobre el tamaño actual de los mismos.

De acuerdo con los datos analizados, si se realizan cargas diarias y almacenan todos los *datamarts* en un solo servidor de bases de datos que posea una capacidad de almacenamiento en disco de 150 GB, se excedería su capacidad en 7.3 años. Así, el Sistema Bodega de Datos no es óptimo ni funcional, debido a que un sistema bodega de datos cualquiera debe por lo menos garantizar el almacenamiento de sus datos en 10 años. Por el contrario, si se implementa un servidor de bases de datos por cada *datamart*, para un total de cuatro servidores, la vida útil del sistema se ve abundantemente superada. Como ejemplo se toma el *datamart* de investigaciones, cuyo peso es de 3,52 GB por año, el cual tendría una vida útil a nivel de almacenamiento de 42,61 años.

Conclusiones

Los diferentes Sistemas de Información usados en la actualidad, se implementan para dar cubrimiento a requerimientos específicos de las organizaciones y sólo se centran

en realizar o resolver ciertos procesos operativos. Esto ocasiona que se tenga una ventaja competitiva en ciertos campos pero no se encuentran integradas con las demás Sistemas de Información para ofrecer una solución integrada.

A partir de lo anterior, se tomó como base el modelo de indicadores del SUE que se analizó, lo que permitió ubicar las fuentes de información y establecer los alcances y las limitaciones y modelo de indicadores a implementar en el Sistema Bodega de Datos SUE. Una vez determinados los elementos anteriormente mencionados, se estableció la estructura o arquitectura que mejor se acoplaba a las necesidades de la institución. Dicha acción consistió en realizar una arquitectura basada en *datamarts* independientes a nivel de inteligencia de negocios porque permite la inclusión, de forma flexible, de nuevos departamentos o áreas de negocio, sin afectar el funcionamiento ni diseño de los demás. Por otro lado, es más fácil implementar la escalabilidad del Sistema Bodega de Datos SUE mediante esta arquitectura y como la institución se encuentra en una etapa de sistematización de sus procesos, este tipo de arquitectura da la flexibilidad necesaria para la inclusión de estos nuevos procesos.

Como producto final se implementó el Sistema Bodega de Datos SUE, el cual como se evidencia en los resultados mostrados, es capaz de capturar la información desde las fuentes identificadas, procesa y limpia los datos, para luego generar las vistas de análisis y los reportes que dan respuesta a los indicadores que plantea el SUE. Todo enmarcado dentro de las limitaciones y alcances definidos.

El Sistema Bodega de Datos SUE es un sistema muy robusto, lo que hace que requiera una infraestructura tecnológica y de comunicaciones que la soporte. Para evidenciar y seleccionar qué tipo de infraestructura tecnológica a usar, se realizó un modelo de rendimiento mediante la aplicación de pruebas de *Stress* a la solución planteada, la cual fue implementada sobre un sistema centralizado. Los resultados del modelo determinaron que en un sistema centralizado el consumo de recursos computacionales (uso de CPU, memoria RAM y tráfico de red) por parte del Sistema Bodega de Datos, eran muy altos y requería un sistema distribuido que redujera el uso de estos recursos, ofreciera alta disponibilidad y tolerara fallos hacia los usuarios finales.

Se implementó un sistema distribuido mediante un clúster con balanceador de cargas el cual reparte las solicitudes de usuarios entre los nodos del clúster mejorando el rendimiento, reduciendo el uso de recursos y generando menor tráfico de red. Ello se traduce para los usuarios en un sistema más estable, mucho más disponible y con mejores tiempos de respuesta, como fue en la sección de resultados.

Sumado a lo anterior a nivel del Sistema Distribuido y en concordancia con los datos, así se realizaran cargas diarias y almacenaran todos los *datamarts* en un solo servidor de bases de datos, el cual posea una capacidad de almacenamiento en disco de 150 GB, se excedería su capacidad en 7.3 años, lo cual para el Sistema Bodega de Datos no es óptimo ni funcional. Esto se debe a que un sistema bodega de datos cualquiera debe por lo menos garantizar el almacenamiento de sus datos de por lo menos 10 años. Por el contrario, si se implementa un servidor de bases de datos por cada *datamart*, para un total de cuatro servidores, la vida útil del sistema será abundantemente superada. Como ejemplo se toma el *datamart* de investigaciones, cuyo peso es de 3,52 GB por año, el cual tendría una vida útil a nivel de almacenamiento de 42,61 años.

Referencias

Aleixos, N., Patalano, S., Contero, M., Company, P., y Vila, C. (2001). Metodología top-down para la modelación CAD avanzada: desarrollo del modelo paramétrico asociativo de un radiador de automóvi. En *Xiii congreso internacional de ingeniería* (p. 1-6).

Cano, J. (2007). *Business intelligence: Competir con información* (1.ª ed.). Banesto Fundación Cultural.

Coulouris, G., Dollimore, J., Kindberg, T., y Blair, G. (2011). *Distributed systems: Concepts and design* (5.ª ed.). Addison Wesley.

Kimball, R., y Ross, M. (2002). *The data warehouse toolkit: The complete guide to dimensional modeling* (2.ª ed.). Wiley.

Moss, L., y Atre, S. (2003). *Business intelligence roadmap: The complete project lifecycle for decision-support applications* (1.ª ed.). Addison-Wesley Professional.

Sistema Universitario Estatal (SUE). (2001). *Indicadores de gestión para las universidades públicas.* On line.