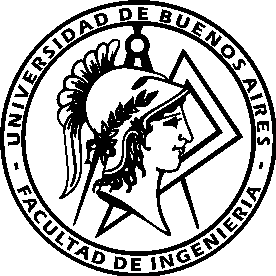
|  |
| --- |
| Universidad de Buenos AiresFacultad de Ingeniería |
| Proyecto de Tesis de Ingeniería Informática |
| Optimización de consultas en un ambiente de soporte de decisiones y el problema de la selección de vistas materializadas |
|  |
| Alumna**:** Tamiozzo, Ayelén Sol  Padrón**:** 90746  Director**:** Dr. Ale, Juan  Fecha: 19/11/2013 |
|  |



|  |
| --- |
|  |

ContenIDOS

[Introducción 3](#_Toc380337774)

[Definición 3](#_Toc380337775)

[Características 3](#_Toc380337776)

[Motivación 3](#_Toc380337777)

[Objetivo de la Tesis 4](#_Toc380337778)

[Metodología de la Tesis 4](#_Toc380337779)

[Estado del Arte 4](#_Toc380337780)

[Cronograma 6](#_Toc380337781)

[Cronograma detallado 7](#_Toc380337782)

[Bibliografía 8](#_Toc380337783)

[Currículum Autora 10](#_Toc380337784)

[Datos personales 10](#_Toc380337785)

[Formación Académica 10](#_Toc380337786)

[Informática 10](#_Toc380337787)

[Idiomas 11](#_Toc380337788)

[Experiencia 11](#_Toc380337789)

[Materias aprobadas 12](#_Toc380337790)

[Plan de Cursada 14](#_Toc380337791)

# Introducción

En la actualidad, el obtener, almacenar y ser capaz de procesar gran cantidad de información es uno de los principales objetivos de muchas aplicaciones. En este contexto, son claves las bases de datos, especialmente los "Data Warehouses", capaces de guardar toda esa información y a través de consultas, procesar y generar informes de gran volumen. Sin embargo, este gran flujo de contenidos provoca que sea crítica la optimización de consultas, para que las mismas tengan un tiempo de respuesta acotado.

En este contexto, una de las decisiones más importantes en el diseño de las bases de datos es la selección de las vistas a materializar. El objetivo es elegir un apropiado conjunto de vistas que minimicen el tiempo total de respuesta de consultas a la vez que satisfagan ciertas restricciones de recursos.

## Definición

La selección de vistas es considerada un problema de optimización que puede ser descripta como una tupla (S,V,M,Q), donde:

* S es el esquema de la base de datos
* V es el conjunto de todas las posibles vistas a elegir
* M es la restricción de recursos a considerar (por ejemplo, cantidad de espacio disponible)
* Q es el conjunto de consultas.

Dadas estas variables, el problema es encontrar en el esquema S un subconjunto V, tal que optimice las consultas de Q mientras que satisfaga los requerimientos M.

## Características

La selección de vistas es uno de los problemas más desafiantes en data warehousing. (Windom, 1995)

La selección de vistas materializadas se define como un problema NP - hard (no polinomial complejo), donde existe una reducción directa derivada del uso de una cobertura mínima del conjunto de datos. (Karloffand & Mihail, 1999)

# Motivación

Hay numerosas motivaciones para la investigación de la selección de vistas a materializar.

En primer lugar, en el contexto actual en el que obtener, almacenar y ser capaz de procesar gran cantidad de información es uno de los principales objetivos de muchas aplicaciones, se vuelve crucial el tiempo de acceso a esta información. Por tal razón, los sistemas comerciales de base de datos desarrollan, con el fin de dar soporte a la administración de las mismas, cada vez herramientas más poderosas, entre las cuales se encuentra aquellas relacionadas con el problema de selección de un apropiado conjunto de vistas a materializar.

Otro motivo para desarrollar esta investigación es el crecimiento de las bases de datos cuyo propósito es la resolución de consultas para el soporte de decisiones en los sistemas *Data Warehouses*, en los cuales los volúmenes de información son enormes y el seleccionar vistas a materializar se convierte en una decisión crucial para su correcto funcionamiento.

Adicionalmente, nuevas aplicaciones de este problema se presentan en la las bases de datos distribuidas, así como también *peer to peer*. (Gribble, Halevy, Ives, Rodrig, & Suciu, 2001)

# Objetivo de la Tesis

La presente tesis tiene como objetivo la investigación y el análisis del problema previamente planteado: como seleccionar las vistas a materializar en un ambiente de soporte de decisiones con el fin de optimizar el tiempo en respuesta de la base de datos.

# Metodología de la Tesis

En la presente tesis, se investigarán algoritmos que se enfoquen en el problema de las vistas a materializar. Se buscará tener una vista global del estado del arte de este tema, entendiendo y analizando cada investigación realizada.

Posteriormente, se seleccionará uno de los algoritmos estudiados a efectos de ampliarlo en lo posible, analizando el orden, performance, funcionamiento, etc.

Habiendo realizado esto, se implementará el algoritmo y finalmente se analizarán posibles optimizaciones / mejoras y se buscará desarrollarlas desde un punto de vista teórico y, de ser posible, también implementarlas.

Finalmente, se probará, analizará y documentará los resultados obtenidos de la comparación del algoritmo original con él propuesto, tanto desde un vista teórico como práctico. También se documentará la investigación del estado del arte previamente realizada.

# Estado del Arte

La selección de vistas se ha estudiado por numerosos autores en varias ocasiones, abordando el problema desde distintos puntos de vista. En el siguiente gráfico se puede ver una clasificación de los métodos de selección estudiados hasta el momento:

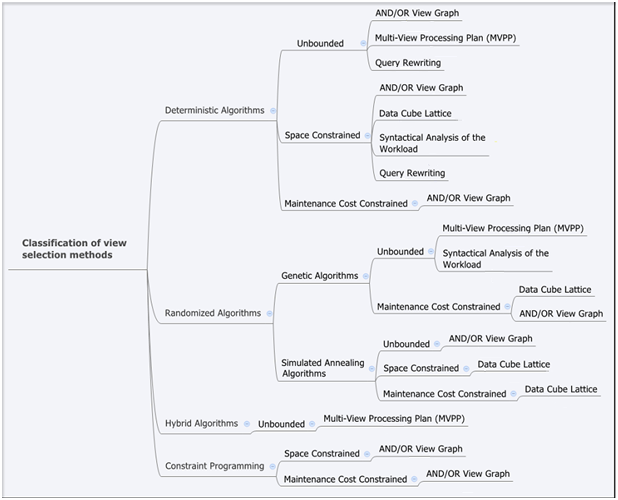


Ilustración 1: Clasificación de métodos de selección de vistas (Mami & Bellahsene, 2012)

Se han clasificado los métodos de acuerdo a distintas dimensiones que caracterizan los algoritmos:

* restricciones de recursos que se consideran durante el proceso de selección
* frameworks que se utilizan para obtener las vistas candidatas

Las heurísticas más conocidas propuestas para la solución del problema son:

* algoritmos determinísticos
* algoritmos aleatorios
* algoritmos híbridos
* programación con restricciones

# Cronograma

Los plazos estipulados para la elaboración de la Tesis se encuentran definidos en el siguiente cronograma:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 |
| Búsqueda de información bilbiográfica |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Análisis de "*A\* heuristic*" |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Implementación de "*A\* heuristic*" |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Desarrollo posibles optimizaciones |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Prueba y análisis de resultados obtenidos |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Documentación de los resultados |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

Las presentes semanas con semanas netas dedicadas al desarrollo de la tesis.

## Cronograma detallado

Dentro del cronograma presentado anteriormente, las horas se dividirán en la siguiente manera:

* 300 horas: búsqueda de información bibliográfica
* 60 horas: análisis del caso de estudio
* 180 horas: implementación y desarrollo de las pruebas
* 180 horas: redacción del trabajo

Total: 720 horas.

# Bibliografía

Aouiche, K., & Darmont, J. (2007). Data Mining-based Materialized View and Index Selection in Data Warehouses.

Asgharzadeh, Z., & Chirkova, R. (2008). Exact and Inexact Methods for Selecting Views and Indexes for OLAP Perfomance Improvement. *EDBT* .

Baril, X., & Bellahsène. (2002). Selection of Materialized Views: a Cost-Based Approach.

Chaudhuri, S. (1998). An Overview of Query Optimization in Relational Systems.

Chaudhuri, S. (1998). AutoAdmin "What-if" Index Analysis Utility. *In SIGMOD Conference* .

Chaudhuri, S. (2000). Automated Selecion of Materialized Views and Indexes for SQL Databases. *26th International Conference on Very Large Database* .

Chauduri, S., & Narasayya, V. (1997). An Efficient, Cost-Driven Index Selection Tool for Microsoft SQL Server. *Proceedings of the 23rd VLDB Conference Athens, Greece* .

Goldstein, J., & Larson, P.-A. (2001). Optimizing Queries Using Materialized Views: A Practical, Scalable Solution. *ACM SIGMOD 2001* .

Gribble, S., Halevy, A., Ives, Z., Rodrig, M., & Suciu, D. (2001). What Can Databases do.

Gupta, H., & Mumick, I. (1999). Selecction of Views to Materialize Under a Maintenance Cost Constraint. *ICDT* .

Gupta, H., & Mumick, I. (2005). Selection of View to Materialize in a Data Warehouse.

Gupta, H., Harinarayan, V., Rajaraman, A., & Ullman, J. (1996). Index Selection for OLAP. *ACM SIGMOD* .

Karloffand, H., & Mihail, M. (1999). On the complexity of the view-selection problem. *PODS* , 167-173.

Liang, W., Wang, H., & Orlowska, M. (2001). Materialized view Selection Under the Maintenance tiem Contraint.

Mami, I., & Bellahsene, Z. (2012). A Survey of View Selection Methods. *SIGMOD* .

Roussopooulos, N. (1986). The Logical Access Path Schema of a Database. *IEEE Transactions on Software engineering, Vol. SE-8, No. 6* .

Theodoratos, D., Ligoudistianos, S., & Sellis, T. (1999). Designing the Global Data Warehouse with SPJ Views. *CAiSE* , 180–194.

Ullman, J., & Garcia-Molina, H. (2008). *Database System: The Complete Book* (2nd ed.). Prentice Hall.

Ullman, J., Rajaraman, A., & Harinarayan, V. (1996). Implementing Data Cubes Efficiently. *SIGMOD* .

Windom, J. (1995). Research problems in data warehousing. *CIKM* , 25-30.

# Currículum Autora

## Datos personales

**Nombre**: Ayelén Sol Tamiozzo **D.N.I.:** 35.148.505

**Domicilio**: Profesor Aguer 5.540 Villa Ballester

**Teléfono**: 4767-6252 15-2-175-3006

**Fecha** de nacimiento: 05/3/1990 **Edad**: 23 años

**Nacionalidad**: argentina **Estado civil**: soltera

## Formación Académica

Instituto Ballester: egresada 2007 en Bachiller Bilingüe con orientación en Ciencias Naturales

Facultad de Ingeniería, Universidad de Buenos Aires: Ingeniería en Informática. Aprobado Ciclo Básico Común y 220 créditos de un total de 248.

## Informática

Sistemas operativos: GNU/Linux (Ubuntu, Debian), Microsoft Windows.

Programación en lenguajes: C/C++, C#, Java, Smalltalk

Herramientas de desarrollo: Javascript, Jquery, Ajax, WPF, ASP.NET, WCF, XML

Base de Datos : SQLServer 2008, MySQL (5.6)

Manejo de programas de control de versión (SVN, Team Foundation Server) y de administración de proyectos (Jira, Assembla)

Metodologías de desarrollo: Extreme Programming, TDD

Programación orientada a objetos, diseño orientado a objetos con UML y patrones de diseño.

## Idiomas

Alemán: Sprachdiplom I (2005), Sprachdiplom II (2007)

Inglés: First Certificate Exam (2007) - Uso diario en el trabajo.

## Experiencia

#### S&P CapitalIQ 05 de marzo 2012 – Actualidad

Cargo: Desarrolladora .Net y SQL Server

*Programación de software financiero para uso interno en la compañía.*

* WPF para interfaz con el usuario
* C# con .Net (4.0) y WCF para la programación de los servicios y modelo del negocio
* Base de datos SQL Server 2008: Desarrollo de scripts de migración, store procedures (con considerable lógica interna) para la obtención de información a mostrar al usuario y jobs para la actualización y sincronización de la información en la base de datos

Tecnologías: C#, WPF, SQL Server, WCF

#### Pines S.A. 13 de junio 2011 – 16 diciembre 2001(Renuncia)

Cargo: Desarrolladora Junior .Net

*Desarrollo de Software bancario*

* Desarrollo y testeo de kits de herramientas para uso interno.
* Diseño e implementación de sitio web intranet para uso bancario.
* Desarrollo de aplicación para la realización de vuelcos de auditoría en bases de dato de gran volumen.

Tecnologías: C#, ASP .Net, Javascript, Jquery, Ajax, SQL

# Materias aprobadas

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Código | Denominación | Nota | Créditos | Cond. | Fecha | Libro | Folio |
| 7540 | ALGORITMOS Y PROGRAMACION I | 8 | 6 - OBL | R | 04-08-2009 | 101 | 82 |
| 6201 | FISICA I A | 6 | 8 - OBL | R | 18-08-2009 | 107 | 178 |
| 6103 | ANALISIS MATEMATICO II A | 5 | 8 - OBL | R | 20-08-2009 | 154 | 96 |
| 6107 | MATEMATICA DISCRETA | 8 | 6 - OPT | R | 29-12-2009 | 156 | 33 |
| 7541 | ALGORITMOS Y PROGRAMACION II | 7 | 6 - OBL | R | 10-02-2010 | 103 | 4 |
| 6108 | ALGEBRA II A | 5 | 8 - OBL | R | 17-02-2010 | 153 | 163 |
| 7801 | IDIOMA INGLES | 8 | 4 - OPT | R | 05-07-2010 | 23 | 22 |
| 6301 | QUIMICA | 9 | 6 - OBL | R | 05-07-2010 | 75 | 69 |
| 6203 | FISICA II A | 6 | 8 - OBL | R | 15-07-2010 | 108 | 118 |
| 7507 | ALGORITMOS Y PROGRAMACION III | 8 | 6 - OBL | R | 03-08-2010 | 104 | 186 |
| 7512 | ANALISIS NUMERICO I | 8 | 6 - OBL | R | 13-12-2010 | 105 | 170 |
| 6602 | LABORATORIO | 7 | 6 - OBL | R | 22-12-2010 | 140 | 4 |
| 6670 | ESTRUCTURA DEL COMPUTADOR | 4 | 6 - OBL | R | 10-02-2011 | 140 | 25 |
| 6109 | PROBABILIDAD Y ESTADISTICA B | 7 | 6 - OBL | R | 17-02-2011 | 159 | 11 |
| 7542 | TALLER DE PROGRAMACION I | 10 | 4 - OBL | R | 29-06-2011 | 106 | 157 |
| 6215 | FISICA III D | 7 | 4 - OBL | R | 20-07-2011 | 109 | 82 |
| 6110 | ANALISIS MATEMATICO III A | 7 | 6 - OBL | R | 21-07-2011 | 157 | 77 |
| 7506 | ORGANIZACION DE DATOS | 4 | 6 - OBL | R | 28-07-2011 | 107 | 80 |
| 7112 | ESTRUCTURA DE LAS ORGANIZACIONES | 6 | 6 - OBL | R | 14-12-2011 | 153 | 84 |
| 7114 | MODELOS Y OPTIMIZACION I | 8 | 6 - OBL | R | 22-12-2011 | 153 | 145 |
| 6620 | ORGANIZACIÓN DE COMPUTADORAS | 6 | 6 - OBL | R | 23-02-2012 | 141 | 197 |
| 7508 | SISTEMAS OPERATIVOS | 7 | 6 - OBL | R | 05-07-2012 | 109 | 59 |
| 7509 | ANALISIS DE LA INFORMACION | 7 | 6 - OBL | R | 10-07-2012 | 109 | 75 |
| 7510 | TECNICAS DE DISEÑO | 7 | 6 - OBL | R | 10-12-2012 | 110 | 128 |
| 7526 | SIMULACION | 10 | 6 - OPT | R | 12-12-2012 | 110 | 158 |
| 7550 | INTRODUCCION A LOS SISTEMAS INTELIGENTES | 8 | 6 - OPT | R | 13-12-2012 | 110 | 168 |
| 7543 | INTRODUCCION A LOS SISTEMAS DISTRIBUIDOS | 7 | 6 - OBL | R | 28-12-2012 | 111 | 28 |
| 7552 | TALLER DE PROGRAMACION II | 8 | 4 - OBL | R | 22-02-2013 | 111 | 128 |
| 7544 | ADM Y CONTROL DE PROYECTOS I | 9 | 6-OBL | R | 01-08-2013 | 0 | 0 |
| 7565 | MANUFACTURA INTEGRADA POR COMP.(CIM) I | 7 | 6- OPT | R | 15-07-2013 | 0 | 0 |
| 7567 | SIST. AUTOM. DE DIAGNÓSTICO Y DETECCIÓN DE FALLAS I | 8 | 6- OPT | R | 29-07-2013 | 0 | 0 |
| 7545 | TALLER DE DESARROLLO DE PROYECTOS I | 9 | 6-OBL | R | 08-08-2013 | 0 | 0 |
| 7515 | BASE DE DATOS | 8 | 6-OBL | R | 15-07-2013 | 0 | 0 |
| 7546 | ADMINISTRACION Y CONTROL DE PROYECTOS 2 | 7 | 6-OBL | R |  | 0 | 0 |
| 7140 | LEGISLACIÓN Y EJERCICIO PROF. ING INF. | 9 | 4-OBL | R |  | 0 | 0 |
| 7547 | TALLER DE DESARROLLO DE PROYECTOS II | 9 | 6-OBL | R |  | 0 | 0 |
| 7113 | INFORMACION DE LAS ORGANIZACIONES | 5 | 6-OBL | R |  | 0 | 0 |

Total Créditos: 220/248 Promedio: 7,27

# Plan de Cursada

|  |  |  |  |
| --- | --- | --- | --- |
| Código | Denominación | Créditos | Fecha |
| 7548 | CALIDAD EN DESAROLLO DE SISTEMAS | 4-OBL | 1 - 2014 |
| 7500 | TESIS | 24-OBL | 2 - 2014 |

Total Créditos: 28