

**GRID COMPUTING PARA MODELOS MATEMATICOS  
E INDEXACION DE INFORMACION NO ESTRUCTURADA**

**ANDRES VERA**

**FUNDACION UNIVERSITARIA SAN MARTIN  
FACULTAD DE INGENIERIA  
PROGRAMA DE INGENIERIA SISTEMAS**

**BOGOTA**

**2010 I**

GRID COMPUTING PARA MODELOS MATEMATICOS E INDEXACION DE  
INFORMACION NO ESTRUCTURADA

T/Monogr.  
670.001  
V4739  
2010  
E11

ANDRES VERA

FUNDACIÓN UNIVERSITARIA SAN MARTÍN  
FACULTAD DE INGENIERÍA  
PROGRAMA DE INGENIERÍA SISTEMAS  
BOGOTÁ  
2010 I

**GRID COMPUTING PARA MODELOS MATEMATICOS E INDEXACION DE  
INFORMACION NO ESTRUCTURADA**

**ANDRES VERA  
COD. 012529  
andresv4444@yahoo.es**

**MONOGRAFÍA DE GRADO**

**ASESOR TÉCNICO  
ING. CRISTIAN BENAVIDEZ**

**FUNDACIÓN UNIVERSITARIA SAN MARTÍN  
FACULTAD DE INGENIERÍA  
PROGRAMA DE INGENIERÍA SISTEMAS  
BOGOTÁ  
2010 I**

**Nota de aceptación**

---

---

---

---

---

---

---

**Ing. Cristian Benavidez  
Asesor**

---

**Nombre Jurado 1  
Jurado 1**

---

**Nombre Jurado 2  
Jurado 2**

## AGRADECIMIENTOS

A Dios, por brindarnos la dicha de la salud y bienestar físico y espiritual

A mi Madre por darme animo en los momentos más difíciles y por su apoyo incondicional.

A mi novia Marisol por darme su comprensión en los momentos en los cuales no pudimos compartir debido al esfuerzo que se requirió para sacar este proyecto

A mis hijos por ser mi fuente de inspiración y la razón por la cual no desistí de este proyecto.

A todos los docentes que hicieron parte de mi formación profesional

## CONTENIDO

	pág.
AGRADECIMIENTOS	4
CONTENIDO	5
LISTA DE TABLAS	11
LISTA DE FIGURAS	12
RESUMEN	16
PROBLEMA	18
2. JUSTIFICACIÓN	19
3. OBJETIVOS	20
3.1 Objetivo General	20
3.2 Objetivos Específicos	20
4. MARCO REFERENCIAL	21
4.1 ANTECEDENTES	21
4.1.1 EGEE (Enabling Grids for E-Science)	21
4.1.2 Teragrid (Teragrid, s.f)	21
4.1.3 EELA (E-Infrastructure shared between Europe and Latin America) (EELA science Grid, 2008)	21
4.1.4 CLGRID (Iniciativa de Grid nacional) (REUNA Red Universitaria Nacional, 2007)	22
4.1.5 Modelos Matemáticos en el Campo de la Biología (Bioinformática y Grid, 2006)	22

4.1.6 Caracterización de Yacimientos (Natfrac Corporación, 2006)	23
4.1.7 Caracterización de Yacimientos (Natfrac Corporación, 2006)	27
4.2 MARCO CONCEPTUAL	29
4.2.1 Optimización (Linares, 2001)	29
4.2.1.1 Métodos de Optimización (Linares, 2001)	30
4.2.1.2 Modelo Matemático (Linares, 2001)	34
4.2.1.3 Etapas en el desarrollo de un modelo (Linares, 2001) (Burgardt, 2005)	34
4.2.1.4 Lenguajes de Modelado Algebraico (Ramo, 2008)	36
4.2.1.5 Lenguajes de Modelado Algebraico (Ramo, 2008)	37
4.2.2 ¿Qué es GRID COMPUTING? (GridGain Cloud Computing, 2009)	40
4.2.2.1 Funcionamiento del Grid	40
4.2.2.2 La Arquitectura Grid (Sánchez & Villafranca, 2005)	41
4.2.2.3 Comparación entre tecnologías distribuidas conocidas y Grid (Domínguez Hernández, 2005)	42
4.2.3 Indexadores de google (GoogleBot, 2010)	42
4.3 MARCO TEÓRICO	43
4.3.1 GLPK (GLPK, 2008)	43
4.3.2 Interface de conectividad entre GLPK y Java (GLPK, s.f.)	43
4.3.3 Ejemplo de un modelo en GLPK (Ceron, 2006)	44
4.3.3.1 Xpress-MP (Gueret, Prins, Sevaux, 2007)	47

4.3.3.2 Usando Xpress-Mosel (Gueret, Prins, Sevaux, 2007)	49
4.3.3.3 Estructura de un modelo Xpress (Gueret, Prins, Sevaux, 2007)	49
4.3.3.4 Ejecutar un modelo Xpress en Java (Dash Optimization, 2006)	50
4.3.3.5 Ejemplo de un programa Java BCL (Fico Dash Optimization, 2008)	50
(Dash Optimization, 2006)	51
4.3.3.6 Referencia de las clases de Java (Dash Optimization, 2006)	51
4.3.4 RCP (Eclipse Plugin Central, 2009) (Mcafeer, Lemieux, 2006)	52
4.3.4.1 Eclipse RCP (Llorens Vilella, López Sacanell, Pardo Invernón, 2006)	52
4.3.4.2 Desarrollo de una aplicación RCP (Coplec, 2008a)	53
4.3.4.3 Elementos de una aplicación RCP (Coplec, 2008b)	63
4.3.4.4 Clases que componen un proyecto RCP (Coplec, 2008b)	66
4.4 Hadoop Distributed File System (Tom White, 2009)	67
4.4.1 Producto de hardware	68
4.4.2 Namenodes y Datanodes	68
4.4.3 GridGain (Gridgain Cloud Computing, 2009)	69
4.4.4 Características principales de GridGain (Gridgain Cloud Computing, 2009)	70
4.4.5 Programación Orientada por aspectos (Reina Quintero, 2000)	71
4.4.6 Arquitectura de Gridgain (Gridgain Cloud Computing, 2009)	72
4.5 Map Reduce (Kleber, 2008)	74
4.6 Estado del Arte	75

4.6.1 GRID COLOMBIA – Proyecto Renata (Castro, Chacón, Díaz, González, Zuluaga, s.f)	75
5. DISEÑO METODOLÓGICO	78
5.1 Open up (Open Unified Process). (Kroll, Macisaac, 2006) (Eclipse, 2009)	78
5.1.1 Características de OpenUP	79
5.1.2 Roles de la metodología openUP (Velzen, 2008)	80
5.1.3 Ciclos de Vida de un Proyecto OpenUP (Velzen, 2008)	81
5.1.4 Iniciación	82
5.2 Elaboración	84
5.2.1 Construcción	85
5.2.2 Transición	86
5.2.3 ¿Por qué OPEN UP y no otra metodología? (Eclipse, 2009) (Kroll, Macisaac, 2006)	87
6. DESARROLLO	91
6.1 Iteración 1	91
6.1.1 Preparación del ambiente de trabajo para la ejecución de la infraestructura de Grid	91
6.2 Iteración 2	92
6.2.1 Desarrollo de objetivos específicos – OpenUp fase construcción	92
6.2.1.1 Objetivo definición de procedimiento de instalación y configuración de una infraestructura de Grid	92

6.2.1.2	Objetivo definición de mecanismo de integración entre Hadoop y GridGain	102
6.2.1.3	Objetivo indexación distribuida de información no estructurada	108
6.2.1.4	Restricciones	109
6.3	Elementos de análisis	109
6.3.1	Clasificación de Información	109
6.3.2	Distribución de archivos a indexar por nodo	110
6.3.3	Almacenamiento de índices	110
6.3.4	Consolidación de índices	111
6.3.5	Librerías Utilizadas	113
6.3.5.1	Hadoop	113
6.3.5.2	Lucene	113
6.3.5.3	GridGain	113
6.3.6	Arquitectura de integración Hadoop y GridGain	114
6.3.7	Diagrama de clases	115
6.4	Iteración 3	118
6.4.1	Objetivo ejecución de modelos matemáticos en Grid	118
6.4.2	Comunicación entre el Plug-in y el Grid	119
6.4.2.1	Diagrama de clases Plug-in RCP desarrollo	121
6.4.2.2	Distribución de los modelos en los nodos del Grid	122

6.4.3 Objetivo modulo RCP que permite recuperar los resultados de los modelos resueltos	126
7. CONCLUSIONES	128
7.1 Desarrollo de objetivos	128
7.2 Inconvenientes	128
7.3 Metodología	129
7.4 Aplicaciones	129
8. RECURSOS	130
8.1 Recurso Humano	130
8.2 Recurso de Hardware	130
8.3 Recurso de Software	130
GLOSARIO	132
BIBLIOGRAFÍA	135

## LISTA DE TABLAS

	pág.
Tabla 1. Método de programación lineal	31
Tabla 2. Método de programación lineal entera mixta	31
Tabla 3. Método de programación cuadrática	31
Tabla 4. Método de programación no lineal	32
Tabla 5. Método de programación multiobjetivo	32
Tabla 6. Métodos meta heurísticos	33
Tabla 7. Otros métodos de optimización	33
Tabla 8. Problemas de programación lineal por tamaños	35
Tabla 9. Elementos de un Modelo Matemático	37
Tabla 10. Comparativo tecnologías distribuidas vs Grid	42
Tabla 11. Micro-Tareas. Iteración 1 Documentación Anteproyecto	88
Tabla 12. Micro-Tareas. Iteración 2 Documentación Anteproyecto	88
Tabla 13. Micro-Tareas. Iteración 2 Documentación RUP	88
Tabla 14. Micro-Tareas. Iteración 3 Documentación RUP	88
Tabla 15. Micro-Tareas. Iteración 1 Documentación RUP	89
Tabla 16. Micro-Tareas. Iteración 1 Diseño de procedimiento de instalación y configuración de infraestructura Grid Computing	89
Tabla 17. Micro-Tareas. Iteración 2 Desarrollo de Requerimientos	90
Tabla 18. Micro-Tareas. Iteración 3 Desarrollo de Requerimientos	90
Tabla 19. Micro-Tareas. Iteración 1 Documentación Monografía	90
Tabla 20. Micro-Tareas. Iteración 2 Fase de Transición	91
Tabla 21. Recurso Humano del proyecto Grid Computing	130
Tabla 22. Recurso Hardware del proyecto Grid Computing	130
Tabla 23. Recurso de Software del proyecto Grid Computing	130

## LISTA DE FIGURAS

	pág.
Figura 1. Caracterización de Yacimientos	23
Figura 2. Simulación de Yacimientos	26
Figura 3. Temperatura de la tierra estable	28
Figura 4. Ecuación diferencial para cambio de temperatura	29
Figura 5. Ecuación para el cálculo de la nueva temperatura	29
Figura 6. Estructura de un Grid Computing	40
Figura 7. Arquitectura Grid	41
Figura 8. Salida del GLPSOL	46
Figura 9. Solución problema de Giapetto: giapetto.sol	47
Figura 10. Árbol del proceso Branch and Bound (B&B)	48
Figura 11. Archivo fuente	51
Figura 12. Arquitectura eclipse RCP	53
Figura 13. Opciones para crear un proyecto	54
Figura 14. Interfaz para ingresar el nombre del proyecto	55
Figura 15. Datos requeridos para generar el Plugin	56
Figura 16. Plantillas de generación del Plugin	57
Figura 17. Adicionar Branding herramientas de aplicaciones RCP	58
Figura 18. Vista general del Plugin	58
Figura 19. Elementos que componen una Aplicación RCP	59
Figura 20. Ejecución del archivo "plugin.xml"	59
Figura 21. Aplicación generada	60
Figura 22. Procedimiento para crear una aplicación	60
Figura 23. Ingreso del nombre de la aplicación	61
Figura 24. Exportar el producto	62
Figura 25. Definir el root directory	62
Figura 26. Directorio y archivos generados	63
Figura 27. Overview: información general	64
Figura 28. Dependencias: Dependencias de un Plugin con otro	64

Figura 29. Runtime: Definir classpath	65
Figura 30. Extensions: Elementos propios de la aplicación	65
Figura 31. Crear extensiones	66
Figura 32. Build. Contenido del archivo jar	66
Figura 33. Arquitectura de hadoop distributed file system	69
Figura 34. Estructura de un programa orientado a aspectos	71
Figura 35. Programación orientada por aspectos	71
Figura 36. Capas de la Arquitectura de GRIDGAIN	72
Figura 37. Esquema de funcionamiento del Map Reduce	75
Figura 38. Capas de la metodología OpenUP	79
Figura 39. Roles de la Metodología OpenUP	80
Figura 40. Fases del Ciclo de Vida de un Proyecto	82
Figura 41. Fase de Inicialización	83
Figura 42. Fase de Elaboración	84
Figura 43. Fase de Construcción	85
Figura 44. Fase de Transición	86
Figura 45. Instalación de Infraestructura de Grid Computing	93
Figura 46. Instalación de Infraestructura de Grid Computing 1	93
Figura 47. Instalación de Infraestructura de Grid Computing 2	94
Figura 48. Instalación y configuración GridGain	94
Figura 49. Directorio GridGain y Hadoop	95
Figura 50. Archivo GridGain descargado a directorio Hadoop	95
Figura 51. Variable entorno GridGain	95
Figura 52. Variable definida para poder aplicar cambios	95
Figura 53. Comando de nodo esclavo	96
Figura 54. Servidor Habilitado SSH	96
Figura 55. Clave SSH	97
Figura 56. Acceso a la maquina	97
Figura 57. Instalación de los binarios de Hadoop	97
Figura 58. Archivo descomprimido	97
Figura 59. Renombrar archivo	98

Figura 60. Configuración de archivo	98
Figura 61. Configuración por medio de archivo Conf/slaves y Conf/masters	99
Figura 62. Formateo del HDFS	99
Figura 63. Sube Hadoop con este comando	99
Figura 64. Deshabilitado el modo seguro	99
Figura 65. Nodos que subieron correctamente	100
Figura 66. Ejecutado comando ls	100
Figura 67. Instalación Gpik con Gipsol	101
Figura 68. Instalación Gpik con todas sus dependencias	101
Figura 69. Arquitectura de Hadoop distributed file system	102
Figura 70. Análisis de la anatomía de búsqueda en HDFS	103
Figura 71. Interfaces	103
Figura 72. Lectura estándar de archivos del HDFS	104
Figura 73. Escritura estándar de archivos del HDFS	104
Figura 74. MapReduce GridGain utilizado	105
Figura 75. Implementación de MapReduce utilizado en Hadoop	105
Figura 76. Como Hadoop corre con Map/reduce.	106
Figura 77. Capas de la arquitectura de GridGain	107
Figura 78. Capas de la arquitectura de GridGain 1	109
Figura 79. Distribución de archivos a indexar por nodo	110
Figura 80. Índices con el nombre de cada máquina que compone el nodo	111
Figura 81. Proceso consolidación de Índices	111
Figura 82. Hadoop y GridGain	113
Figura 83. Librerías lucene	113
Figura 84. Librerías GridGain	113
Figura 85. Arquitecturas de Grid: GridGain y Hadoop distributed file system	114
Figura 86. Integración nodos del Grid: Nodos GridGain y Nodos Data Node	115
Figura 87. Diagrama de clases indexación en Grid	115
Figura 88. Diagrama de flujo	116
Figura 89. Integración Plug-in RCP con el Grid	119
Figura 90. Archivo de configuración plugin.xml	120

Figura 91. Dependencias Plug-in RCP	120
Figura 92. Diagrama de clases Plug-in RCP desarrollo	121
Figura 93. Código fuente que distribuye los modelos	122
Figura 94. Diagrama de clases	122
Figura 95. Diagrama de clases flujo de envío de modelos matemáticos al Grid	123
Figura 96. Perspectiva eclipse: Grid modelos matemáticos RCP	124
Figura 97. Configuración RCP: Opciones de configuración controlador GridGain	124
Figura 98. Cargue de modelos: cargue de modelo matemático en lenguaje algebraico Mathpro para GLPK	125
Figura 99. Cargue de modelos: selección de modelos matemáticos a cargar	125
Figura 100. Modulo al Plugins RCP	126
Figura 101. Lista de modelos	126
Figura 102. Visualización modelo resuelto: visualización individual de modelo matemático resuelto	127

## RESUMEN

El proyecto de grado "Grid Computing para modelos matemáticos e indexación de información no estructurada" tiene como objetivo hacer uso de una infraestructura de Grid Computing desplegada, para la solución de problemas computacionales que requieren alta capacidad computacional como lo son la solución de modelos matemáticos de programación lineal, y la indexación de información no estructurada.

Para el desarrollo de este proyecto se baso en el uso de las herramientas:

- **GridGain:** Software de Grid con el cual se logra la distribución de trabajo entre los diferentes nodos de la infraestructura de Grid.
- **Hadoop Distributed File System:** Software con el cual se logra el almacenamiento distribuido.
- **Lucene:** Software con el que se logra realizar la indexación de información no estructurada.
- **Eclipse RCP:** Utilidad de eclipse con la que se logra crear parte de la interfaz de usuario