

List of Figures

1-1	Data cube representing a data warehouse.	4
1-2	A lattice of cuboids	5
1-3	Representation of hierarchies in Data cube	6
1-4	Analytical processing in MapReduce framework.	12
2-1	A lattice structure for 3 attributes	21
2-2	An Expression AND DAG	22
2-3	An Expression AND-OR DAG	22
2-4	An MVPP graph	25
2-5	Relative costs of Heuristic, Evolutionary and Simulated Annealing algorithm in view selection using query processing plan graph representation.	43
2-6	Comparison of GA, PSO and MA based view materialization models with respect to total query processing costs vs. space used by materialized views.	43
3-1	An example MVPP graph using TPC-H benchmark data warehouse	51
3-2	Randomly generated 242 solutions	69
3-3	Distribution of costs by obtained non-dominated solutions after 40 iterations	69
3-4	Distribution of significant representative solutions in objective space	70
3-5	Objective function values of non-dominating solutions	71
4-1	Test-bed for selecting non-dominated solution sets of materialized views	89

List of Figures

4-2	Processing MapReduce cost (in Seconds) by NSGA-II and MODE-BE generated non-dominated solutions.	93
4-3	Materialized view maintenance MapReduce cost (in Seconds) by NSGA-II and MODE-BE generated non-dominated solutions.	93
4-4	Number of views in solution sets for materializing.	93
4-5	Space requirements by solution sets of views for materializing.	94
4-6	Objective functions' values by MODE-BE generated non-dominating solutions.	94
4-7	Objective functions' values by NSGA-II generated non-dominating solutions.	94
5-1	Number of dominating solutions in estimated Pareto front	105
5-2	Domination during initializing an archive of non dominated solutions.	106
5-3	Amount of domination	107
5-4	Clustering to reduce solution population while minimizing objective functions f_1 and f_2	109
5-5	Big data query responding	111
5-6	Restricting solution population in AMOSA-MVS	114
5-7	Purity	127
5-8	Convergence metric (γ)	128
5-9	Minimal spacing between solutions on estimated true Pareto front .	128

List of Tables

2.1	Representations used in view selection algorithms and associated issues	27
2.2	Stochastic algorithm based materialized view selection techniques and associated issues.	41
2.2	Stochastic algorithm based materialized view selection techniques and associated issues.	42
3.1	Performances of multi-objective DE and NSGA-II with respect to DTLZ test problems.	48
3.2	Convergence metric γ of solutions produced by NSGA-II and MODE-BE in different test problems.	66
3.3	Base tables used in our experimental MVPP	68
3.4	Views generated in our experimental MVPP	68
3.5	MODE-BE generated solutions and Convergence metric	73
4.1	Considered frequent HiveQL queries and constituent views	95
4.2	Query responding MapReduce costs of selected queries	96
4.3	Processing and maintenance MapReduce costs and space requirements of candidate views	97
5.1	Convergence (γ)	121
5.6	MODE-BE generated solutions' objective function values considering 109 number of queries and 51 views. Number of initial solutions = 2487.	128
5.10	NSGA-II generated solutions' objective function values considering 109 number of queries and 51 views. Number of initial solutions = 5015	130

5.2	AMOS-A-MVS generated solutions' objective function values considering 109 number of queries and 51 views. Number of initial solutions considered= 3975.	132
5.3	AMOS-A-MVS generated solutions' objective function values considering 60 number of queries and 51 views. Number of initial solutions considered= 3975.	133
5.4	AMOS-A-MVS generated solutions' objective function values considering 50 number of queries and 51 views. Number of initial solutions considered= 3975.	133
5.5	AMOS-A-MVS generated solutions' objective function values considering 20 number of queries and 25 views. Number of initial solutions considered= 4000.	134
5.7	MODE-BE generated solutions' objective function values considering 60 number of queries and 51 views. Number of initial solutions considered= 2545.	135
5.8	MODE-BE generated solutions' objective function values considering 50 number of queries and 51 views. Number of initial solutions considered=2520	136
5.9	MODE-BE generated solutions' objective function values considering 20 number of queries and 25 views. Number of initial solutions considered=2899	136
5.11	NSGA-II generated solutions' objective function values considering 60 number of queries and 51 views. Number of initial solutions considered= 5040.	137
5.12	NSGA-II generated solutions' objective function values considering 50 number of queries and 51 views. Number of initial solutions considered= 5163.	138
5.13	NSGA-II generated solutions' objective function values considering 20 number of queries and 25 views. Number of initial solutions considered=4888	138
5.14	Purity	139
5.15	Minimal Spacing	139

List of Algorithms

1	The HRU Greedy algorithm	29
2	View selection using optimal query plan	33
3	Simulated annealing for selection of views to materialize	34
4	Multi-objective DE using Binary Encoded Data for selecting views to materialize in data warehouse	61
5	Selecting elite N solutions by NSGA-II based non-dominated sorting and SMC based diversity in solution space in Multi-objective DE using Binary Encoded Data	63
6	View selection for materializing by Multi-objective Differential Evolution using Binary Encoded Data in Big data based data warehouse.	84
7	View selection for materializing by Multi-objective Differential Evolution using Binary Encoded Data in Big data based data warehouse- (continued from previous page).	85
8	Initialization of solution population <i>Archive</i>	113
9	Initialization of solution population <i>Archive</i> - (continued)	114
10	Archived Multi-Objective Simulated Annealing for Materialized View Selection (AMOSAMVS)	117
11	<i>Continued- second page</i> - Archived Multi-Objective Simulated Annealing for Materialized View Selection (AMOSAMVS).	118
12	<i>Continued- third page</i> - Archived Multi-Objective Simulated Annealing for Materialized View Selection (AMOSAMVS).	119

Glossary of Terms

ACA	Ant Colony Algorithm
AMOS	Archived Multi-objective Simulated Annealing
AMOS-MVS	Archived Multi-objective Simulated Annealing for Materialized View Selection
AST	Abstract syntax tree
CPU	Central Processing Unit
CSA	Clonal Selection Algorithm
DAG	Directed Acyclic Graph
DE	Differential Evolution
DEMO ^{NS-II}	Differential Evolution based variants of NSGA-II
DFS	Distributed File System
DIMMQ	Discardable In-Memory Materialized Query
DTLZ	K. Deb, L. Thiele, M. Laumanns and E. Zitzler defined test problems for multi-objective optimization techniques.
EA	Evolutionary Algorithm
GA	Genetic Algorithm
GDE3	Generalized Differential Evolution - 3
HDFS	Hadoop Distributed File System
HDP	Hortonworks Data Platform
HRU	Venky Harinarayan, Anand Rajaraman and Jeffrey D. Ullman, Stanford University
MA	Memetic Algorithm
MB	Mega Byte
MODE-BE	Multi-Objective Differential Evolution for Binary Encoded data
MOEA	Multi-Objective Evolutionary Algorithm
MOGA	Multi-Objective Genetic Algorithm
MOO	Multi-Objective Optimization
MOSA	Multi-objective Simulated Annealing
MVPP	Multiple View Processing Plan
NPGA	Niched Pareto Genetic Algorithm
NP-hard	Non-deterministic polynomial-time hard
NSGA-II	Non-dominated Sorting Genetic Algorithm-II
OLAP	On-Line Analytical Processing
OPTICS	Ordering Points to Identify Clustering

	Structure
PB	Petabyte
PGA	Polynomial Greedy Algorithm
PSA	Parallel Simulated Annealing
PSO	Particle Swarm Optimization
RDBMS	Relational Database Management System
RDD	Resilient Distributed Data-set (Spark's)
SA	Simulated Annealing
SMC	Simple Matching Coefficient
SQL	Structured Query Language
SSD	Solid-state drives
TB	Terabyte
TPC	Transaction processing Performance Council
TPC-H	Transaction processing Performance Council (Benchmark version H)
ZDT	Zitzler, Deb and Thiele's test problem for evaluating Evolutionary Multi-objective optimization techniques

Symbols and Notations

\mathbb{B}	Set of all truth values
$\mathbb{E}(S)$	The set of all equivalence relations over a given set S
Δdom	Amount of domination between solutions
$\Delta dom_{average}$	Average domination between solutions
Δdom_{min}	Minimum amount of domination between solutions
δE	Energy difference between solutions generated in Simulated Annealing algorithm
\equiv	Equivalent
γ	Convergence metric of multi-objective optimization technique
Γ	A real valued constant
μMax	Mean of maximum distances between each solution vectors w.r.t all other solution vectors obtained
$\not\prec$	Does not dominate
\prec	Dominates
σMax	Standard deviation of maximum distances between each solution vectors w.r.t all other solution vectors obtained
\sim	A relation
\triangleq	Defined to be equal to
Ξ	The set of <i>formae</i> or equivalence classes
T	System temperature in Simulated Annealing process
A	Availability of memory space for materializing views
$A_{V'}$	Space required for materializing the set of sub-queries V'
A_M	Space required for materializing the set of views M
$B(v, S)$	Total benefit of materializing view v if S is the set of view selected at an iteration of HRU-greedy algorithm
$C(v)$	Cost of processing view v
$C_a^q(v)$	Cost of accessing query q when view or vertex v of an MVPP graph is materialized
C_M^Q	The total cost of responding queries Q when the set of views M is materialized
$C_m^r(v)$	Cost of maintaining a view or vertex v due to change in the base relation r
$C_{total}(v)$	Total processing cost of view v in an MVPP graph.
$C_{V'}^Q$	The total MapReduce cost of processing set of

	queries Q when a set of sub-queries or sub-expressions as views V' is materialized
CR	Real valued cross-over probability constant in the range[0,1]
d	Number of dimensions of a data cube
D_Ψ	The set of different parameters between equivalence relations using basis Ψ .
$DE/x/y/z$	Version of Differential Evolution with x vector to be mutated, y specifying whether <i>random</i> or <i>best</i> and z specifies cross-over scheme
F	Real valued amplification vector in the range [0,1] in DE
$\tilde{\mathcal{F}}$	The current estimate of Pareto front
\mathcal{F}	Pareto front
f_q	Frequency of a query q
f_u	Updating frequency of base relations of an MVPP graph
f_v	Frequency of queries on view v of an AND-OR view graph
g	Generation number in an evolutionary process
g_{max}	Maximum number of generations in an evolutionary process
g_v	Updating frequency of a view v of an AND-OR view graph
L_i	Number of levels of hierarchies associated with dimension i of data warehouse
\mathbf{M}	Objectives of a multi-objective optimization problem
M_{de}	Formae basis based DE mutation operator template
\mathbb{P}	Probability (function value)
\mathcal{P}	Set of Pareto optimal solutions
$Q(v, M)$	Cost of answering sub-query v of an AND-OR view graph when set of views M is materialized
$Q_G(M)$	Total query processing cost of MVPP graph G , if the set of views M is materialized
$R(v)$	Resultant relation corresponding to vertex v in an MVPP graph
R_v	Cost incurred in reading the materialized view v of an AND-OR view graph
$randI()$	Random integer index generator
$randR()$	Evaluation of a uniform random number generator in the range [0,1]
\mathbf{S}	Set of vectors
T_q	Query processing tree for query q
$\tau(G, M)$	Processing cost of AND-OR view graph G when the set of views M is materialized
$U(M)$	The total updating cost of set of materialized views M
$U_G(M)$	Updating cost of MVPP graph G when set of views M is materialized
$UC(v, M)$	Maintenance cost of a sub-query or view v when the set of views M is materialized