Automating Index Selection Using Constraint Programming

Optimization Days 2023

Philippe Olivier[1]

[1]pganalyze, California, USA
philippe.olivier@polymtl.ca

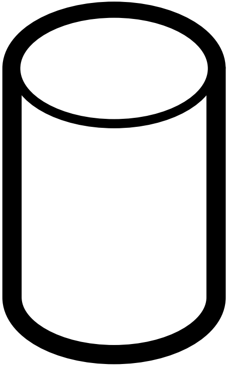
30 May 2023

Outline

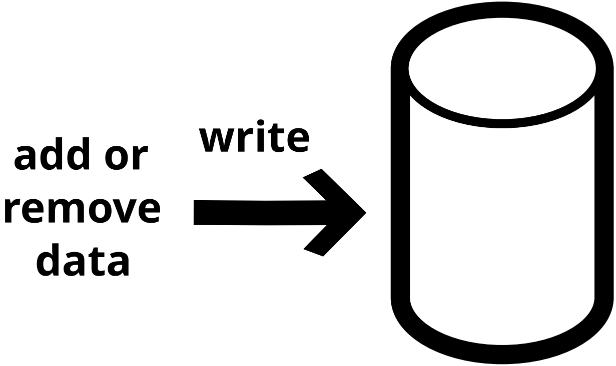
Outline

1. Introduction
2. Index Selection Problem
3. Basic CP and MIP Formulations
4. Multi-Objective CP Model
5. Final Thoughts

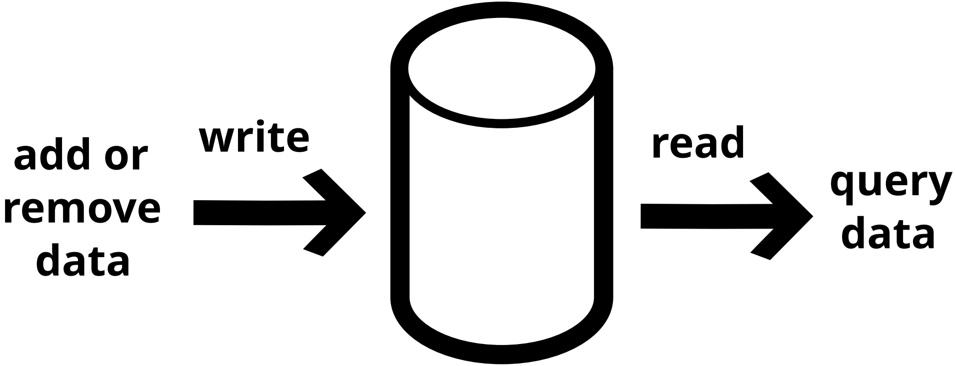
A Database



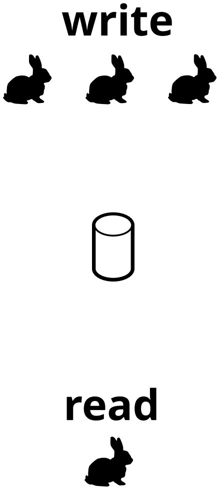
A Database



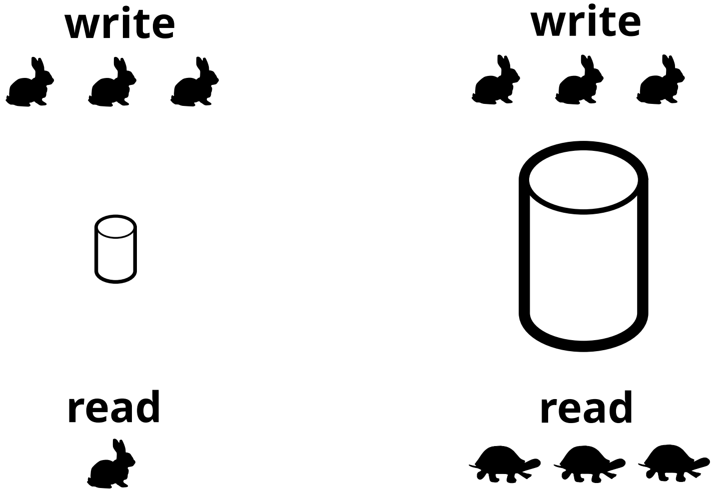
A Database



The Difficulty

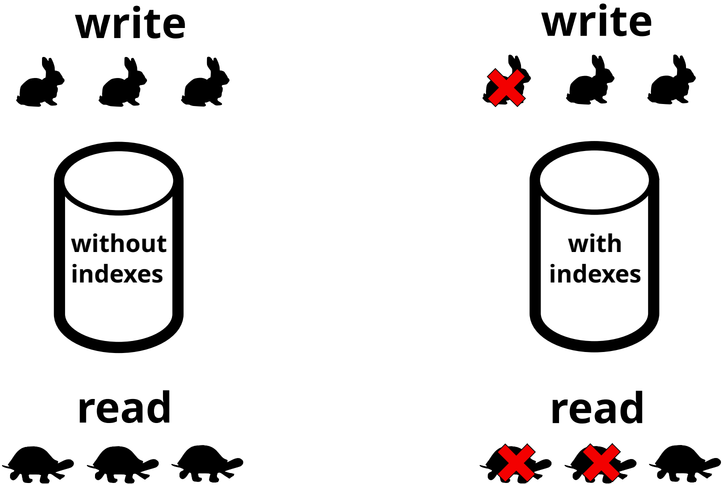


The Difficulty



Adding Indexes

Adding Indexes



Outline

Index Selection Problem

$\mathcal{S}$: Ordered set of m scans

$\mathcal{I}$: Ordered set of n indexes

C: $n \times m$ cost matrix with $\mathrm{c_{ij}}$ the cost of scan j offered by index i

r_j: Default cost of scan j

b_i: Budget cost of using index i (storage, writes, etc)

B: Allowed budget for indexes

Objective: Minimize the costs of the scans

Constraints: Don't exceed the budget B

Which subset of indexes offers the best "performance" for a given "budget"?

Visually

	r →	9	9	9
b ↓		S_1	S_2	S_3
3	I_1	4	3	
3	I_2		3	4
1	I_3	8		5
1	I_4	7	2	8

$B = 5$

Visually

b ↓	r →	9	9	9
		S_1	S_2	S_3
3	I_1	4	3	
3	I_2		3	4
1	I_3	8		5
1	I_4	7	2	8

B = 5

Visually

b ↓	r →	9	9	9
		S_1	S_2	S_3
3	I_1	4	3	
3	I_2		3	4
1	I_3	8		5
1	I_4	7	2	8

B = 5

Visually

	r →	9	9	9
b ↓		S_1	S_2	S_3
3	I_1	4	3	
3	I_2		3	4
1	I_3	8		5
1	I_4	7	2	8

$B = 5$

Outline

Constraint Programming (CP)

Constraint Satisfaction Problem (CSP)

CSP $= \langle X, D, C \rangle$

- X a set of variables
- D the domains (ranges of values) of the variables
- C a set of constraints

Assign values from D to variables in X such that C is satisfied.

Formulations

Constraint Programming

$$\min \sum_{j \in \mathcal{S}} s_j$$

$$s_j \in \left\{ \min_{i \in \mathcal{I}} \{c_{ij}\}, r_j \right\}, \forall j \in \mathcal{S}$$

$$s_j = \min_{i \in \mathcal{I}} \{x_i c_{ij} + r_j(1 - x_i)\}, \forall j \in \mathcal{S}$$

$$\sum_{i \in \mathcal{I}} x_i b_i \leq B$$

$$x_i \in \{0, 1\}, \forall i \in \mathcal{I}$$

Mixed-Integer Programming

$$\begin{aligned}
\min \quad & \sum_{j \in \mathcal{S}} s_j \\
\text{s.t.} \quad & s_j = \sum_{i \in \mathcal{I}} u_{ij} c_{ij} + \\
& \qquad r_j\Big(1 - \sum_{i \in \mathcal{I}} u_{ij}\Big), \forall j \in \mathcal{S} \\
& u_{ij} \leq x_i && \forall i \in \mathcal{I}, \forall j \in \mathcal{S} \\
& \sum_{i \in \mathcal{I}} u_{ij} \leq 1 && \forall j \in \mathcal{S} \\
& \sum_{i \in \mathcal{I}} x_i b_i \leq B \\
& u_{ij} \in \{0, 1\} && \forall i \in \mathcal{I}, \forall j \in \mathcal{S} \\
& x_i \in \{0, 1\} && \forall i \in \mathcal{I}
\end{aligned}$$

Formulations

Constraint Programming

$$\min \sum_{j \in \mathcal{S}} s_j$$

$$s_j \in \left\{ \min_{i \in \mathcal{I}} \{c_{ij}\}, r_j \right\}, \forall j \in \mathcal{S}$$

$$s_j = \min_{i \in \mathcal{I}} \{x_i c_{ij} + r_j(1 - x_i)\}, \forall j \in \mathcal{S}$$

$$\sum_{i \in \mathcal{I}} x_i b_i \leq B$$

$$x_i \in \{0, 1\}, \forall i \in \mathcal{I}$$

Mixed-Integer Programming

$$\begin{aligned}
\min \quad & \sum_{j \in \mathcal{S}} s_j \\
\text{s.t.} \quad & s_j = \sum_{i \in \mathcal{I}} u_{ij} c_{ij} + \\
& \qquad r_j(1 - \sum_{i \in \mathcal{I}} u_{ij}), \forall j \in \mathcal{S} \\
& u_{ij} \leq x_i & \forall i \in \mathcal{I}, \forall j \in \mathcal{S} \\
& \sum_{i \in \mathcal{I}} u_{ij} \leq 1 & \forall j \in \mathcal{S} \\
& \sum_{i \in \mathcal{I}} x_i b_i \leq B \\
& u_{ij} \in \{0, 1\} & \forall i \in \mathcal{I}, \forall j \in \mathcal{S} \\
& x_i \in \{0, 1\} & \forall i \in \mathcal{I}
\end{aligned}$$

Formulations

Constraint Programming

$$\min \sum_{j \in \mathcal{S}} s_j$$

$$s_j \in \left\{ \min_{i \in \mathcal{I}} \{c_{ij}\}, r_j \right\}, \forall j \in \mathcal{S}$$

$$s_j = \min_{i \in \mathcal{I}} \{x_i c_{ij} + r_j(1 - x_i)\}, \forall j \in \mathcal{S}$$

$$\sum_{i \in \mathcal{I}} x_i b_i \leq B$$

$$x_i \in \{0, 1\}, \forall i \in \mathcal{I}$$

Mixed-Integer Programming

$$\begin{aligned}
\min \quad & \sum_{j \in \mathcal{S}} s_j \\
\text{s.t.} \quad & s_j = \sum_{i \in \mathcal{I}} u_{ij} c_{ij} + \\
& \quad r_j(1 - \sum_{i \in \mathcal{I}} u_{ij}), \forall j \in \mathcal{S} \\
& u_{ij} \leq x_i \qquad \forall i \in \mathcal{I}, \forall j \in \mathcal{S} \\
& \sum_{i \in \mathcal{I}} u_{ij} \leq 1 \qquad \forall j \in \mathcal{S} \\
& \sum_{i \in \mathcal{I}} x_i b_i \leq B \\
& u_{ij} \in \{0, 1\} \qquad \forall i \in \mathcal{I}, \forall j \in \mathcal{S} \\
& x_i \in \{0, 1\} \qquad \forall i \in \mathcal{I}
\end{aligned}$$

Formulations

Constraint Programming

$$\min \sum_{j \in \mathcal{S}} s_j$$

$$s_j \in \left\{ \min_{i \in \mathcal{I}} \{c_{ij}\}, r_j \right\}, \forall j \in \mathcal{S}$$

$$s_j = \min_{i \in \mathcal{I}} \{x_i c_{ij} + r_j(1 - x_i)\}, \forall j \in \mathcal{S}$$

$$\sum_{i \in \mathcal{I}} x_i b_i \leq B$$

$$x_i \in \{0, 1\}, \forall i \in \mathcal{I}$$

Mixed-Integer Programming

$$\begin{aligned}
\min \quad & \sum_{j \in \mathcal{S}} s_j \\
\text{s.t.} \quad & s_j = \sum_{i \in \mathcal{I}} u_{ij} c_{ij} + \\
& \quad r_j \Big(1 - \sum_{i \in \mathcal{I}} u_{ij}\Big), \forall j \in \mathcal{S} \\
& u_{ij} \leq x_i & \forall i \in \mathcal{I}, \forall j \in \mathcal{S} \\
& \sum_{i \in \mathcal{I}} u_{ij} \leq 1 & \forall j \in \mathcal{S} \\
& \sum_{i \in \mathcal{I}} x_i b_i \leq B \\
& u_{ij} \in \{0, 1\} & \forall i \in \mathcal{I}, \forall j \in \mathcal{S} \\
& x_i \in \{0, 1\} & \forall i \in \mathcal{I}
\end{aligned}$$

Formulations

Constraint Programming

$$\min \sum_{j \in \mathcal{S}} s_j$$

$$s_j \in \left\{ \min_{i \in \mathcal{I}} \{c_{ij}\}, r_j \right\}, \forall j \in \mathcal{S}$$

$$s_j = \min_{i \in \mathcal{I}} \{x_i c_{ij} + r_j(1 - x_i)\}, \forall j \in \mathcal{S}$$

$$\sum_{i \in \mathcal{I}} x_i b_i \leq B$$

$$x_i \in \{0, 1\}, \forall i \in \mathcal{I}$$

Mixed-Integer Programming

$$\begin{aligned}
\min \quad & \sum_{j \in \mathcal{S}} s_j \\
\text{s.t.} \quad & s_j = \sum_{i \in \mathcal{I}} u_{ij} c_{ij} + \\
& \quad r_j(1 - \sum_{i \in \mathcal{I}} u_{ij}), \forall j \in \mathcal{S} \\
& u_{ij} \leq x_i & \forall i \in \mathcal{I}, \forall j \in \mathcal{S} \\
& \sum_{i \in \mathcal{I}} u_{ij} \leq 1 & \forall j \in \mathcal{S} \\
& \sum_{i \in \mathcal{I}} x_i b_i \leq B \\
& u_{ij} \in \{0, 1\} & \forall i \in \mathcal{I}, \forall j \in \mathcal{S} \\
& x_i \in \{0, 1\} & \forall i \in \mathcal{I}
\end{aligned}$$

Formulations

Constraint Programming

$$\min \sum_{j \in \mathcal{S}} s_j$$

$$s_j \in \left\{ \min_{i \in \mathcal{I}} \{c_{ij}\}, r_j \right\}, \forall j \in \mathcal{S}$$

$$s_j = \min_{i \in \mathcal{I}} \left\{ x_i c_{ij} + r_j(1 - x_i) \right\}, \forall j \in \mathcal{S}$$

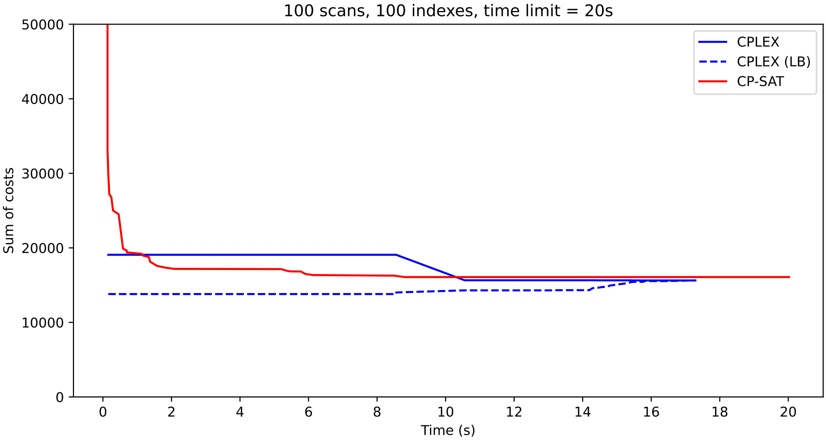
$$\sum_{i \in \mathcal{I}} x_i b_i \leq B$$

$$x_i \in \{0, 1\}, \forall i \in \mathcal{I}$$

Mixed-Integer Programming

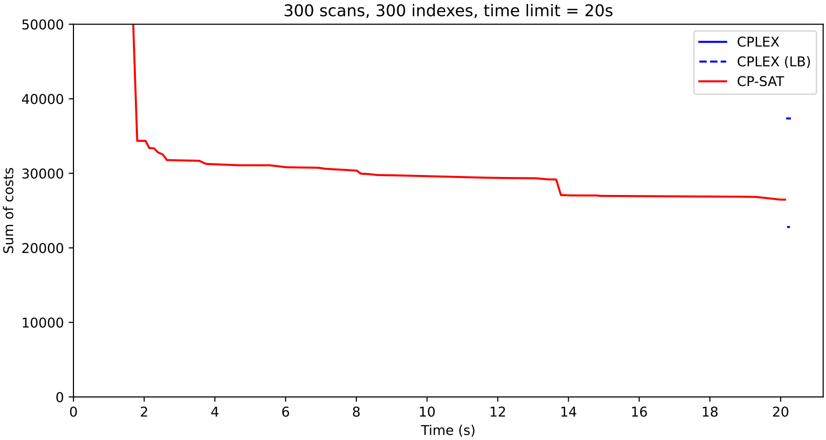
$$\begin{aligned}
\min \quad & \sum_{j \in \mathcal{S}} s_j \\
\text{s.t.} \quad & s_j = \sum_{i \in \mathcal{I}} u_{ij} c_{ij} + \\
& \qquad r_j(1 - \sum_{i \in \mathcal{I}} u_{ij}), \forall j \in \mathcal{S} \\
& u_{ij} \leq x_i && \forall i \in \mathcal{I}, \forall j \in \mathcal{S} \\
& \sum_{i \in \mathcal{I}} u_{ij} \leq 1 && \forall j \in \mathcal{S} \\
& \sum_{i \in \mathcal{I}} x_i b_i \leq B \\
& u_{ij} \in \{0, 1\} && \forall i \in \mathcal{I}, \forall j \in \mathcal{S} \\
& x_i \in \{0, 1\} && \forall i \in \mathcal{I}
\end{aligned}$$

Results

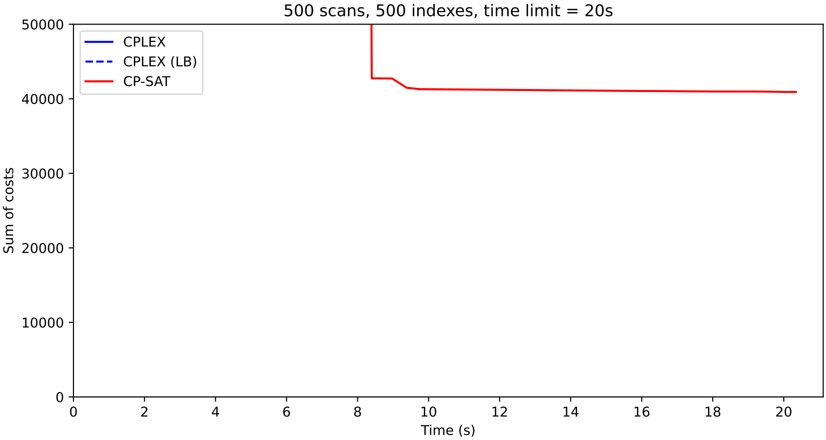


100 scans, 100 indexes, time limit = 20s
Sum of costs
Time (s)
CPLEX
CPLEX (LB)
CP-SAT
0
10000
20000
30000
40000
50000
0
2
4
6
8
10
12
14
16
18
20

Results

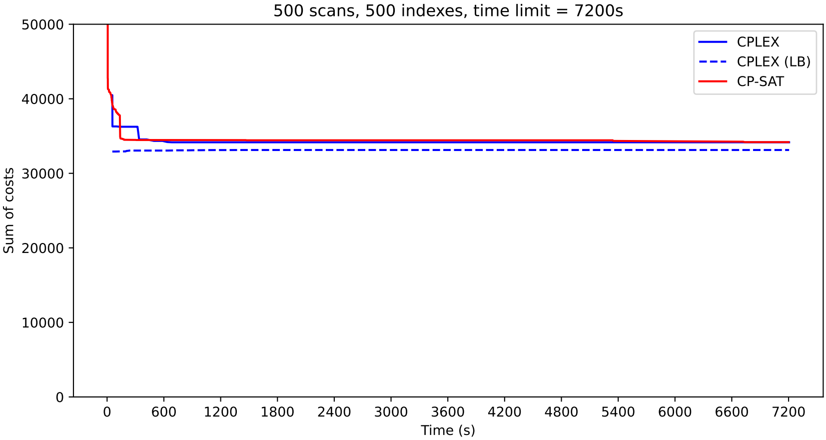

300 scans, 300 indexes, time limit = 20s
50000
40000
30000
20000
10000
0
Sum of costs
0
2
4
6
8
10
12
14
16
18
20
Time (s)
CPLEX
CPLEX (LB)
CP-SAT

Results



500 scans, 500 indexes, time limit = 20s
CPLEX
CPLEX (LB)
CP-SAT
Sum of costs
Time (s)
0
10000
20000
30000
40000
50000
0
2
4
6
8
10
12
14
16
18
20

Results



500 scans, 500 indexes, time limit = 7200s
CPLEX
CPLEX (LB)
CP-SAT
Sum of costs
Time (s)
0
10000
20000
30000
40000
50000
0
600
1200
1800
2400
3000
3600
4200
4800
5400
6000
6600
7200

Advantages and Drawbacks

- Quick and good solutions: **CP-SAT**
- Robustness: **CP-SAT**
- Optimality guarantees: **CPLEX**
- Price: **CP-SAT**

Outline

Objectives and Constraints

- Minimize total scan cost/impact
- Maximize coverage
- Minimize index overhead (storage, writes, etc)
- Minimize the number of indexes
- Minimize worst cost/impact
- Maximum number of indexes/overhead (constraints)
- And more

Hierarchical Optimization Method (Waltz, 1967)

Choose a tolerance value for each objective

For every objective, in lexicographical order:

1. Solve the problem (and find objective value X)
2. New constraint: This objective cannot be worse than X ($\pm$ tolerance) in subsequent steps

Hierarchical Optimization Method (Waltz, 1967)

Choose a tolerance value for each objective

For every objective, in lexicographical order:

1. Solve the problem (and find objective value X)
2. New constraint: This objective cannot be worse than X ($\pm$ tolerance) in subsequent steps

"I want to be within 90% of whatever the lowest possible costs are. How can I achieve this with the fewest indexes?"

Outline

Real-Life Considerations

- Robustness
- Time
- Money

Future

- PGCon 2023 (Ottawa)
- Closed-source/open-source
- Try it: `github.com/pganalyze/pgcon2023`