Romain Perriot
Jérémy Pfeifer
Laurent d'Orazio
Bruno Bachelet
Sandro Bimonte
Jérôme Darmont

Cost Models for Materialized View Selection in the Cloud

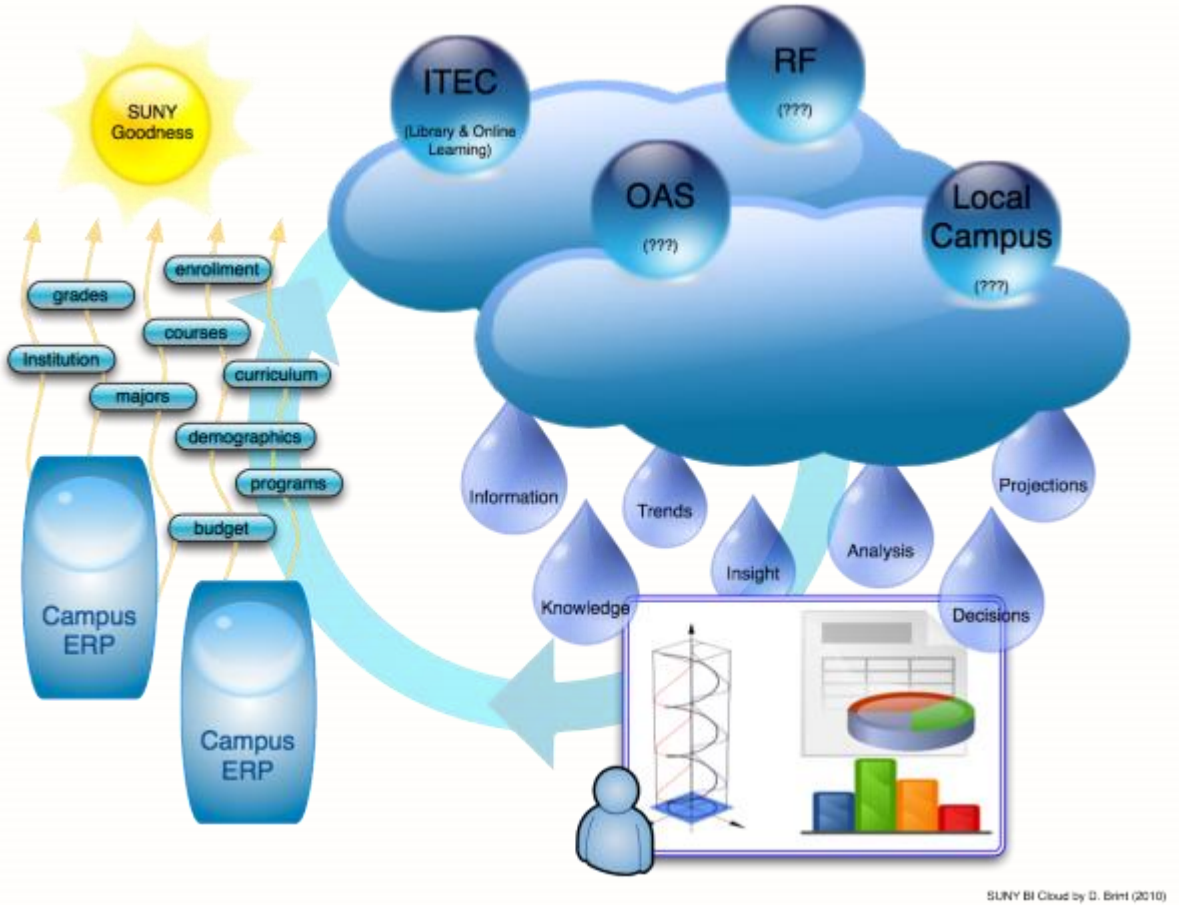
Application to Amazon EC2 and S3 Services

LIMOS

Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture
irstea

eric
Entrepôts, Représentation & Ingénierie des Connaissances

BI in the cloud(s)



danbrint.wordpress.com

Pay more to earn more… performance

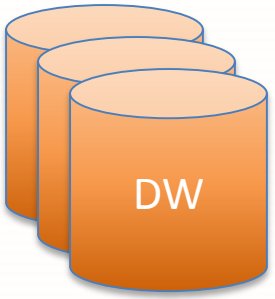
Cloud computing

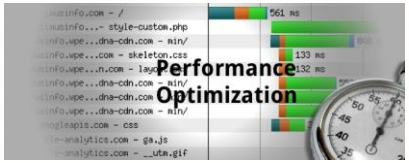
from an economic point of view

Elasticity

Pay-as-you-go

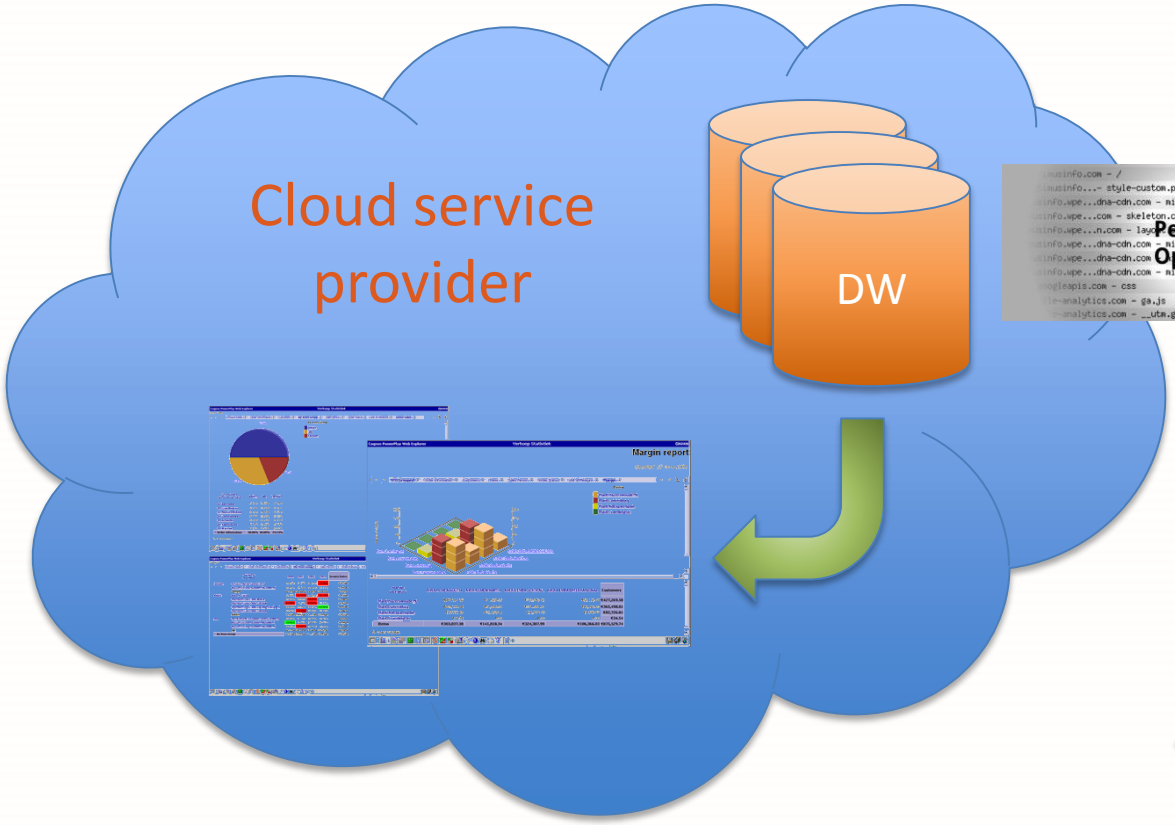
Let's be pragmatic!

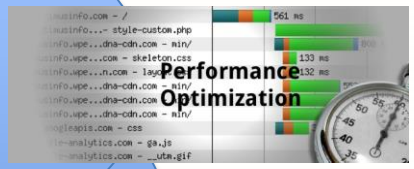




- Indexes
- Caches
- Materialized views
- Fragmentation
- …

Let's be pragmatic!



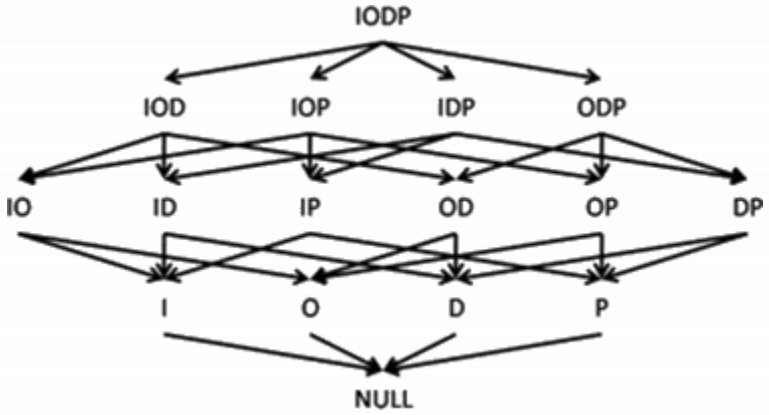


- Indexes
- Caches
- Materialized views
- Fragmentation
- …

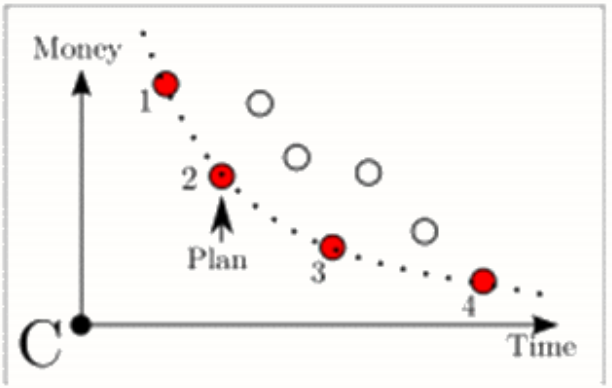
$$Cost_{global} = Cost_{transfer} + Cost_{cpu} + Cost_{storage}$$

Problems and contributions

Selection of views to materialize

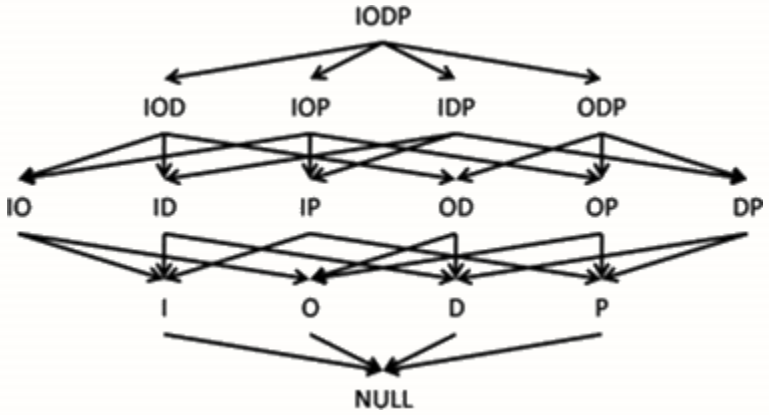


Multicriteria optimization

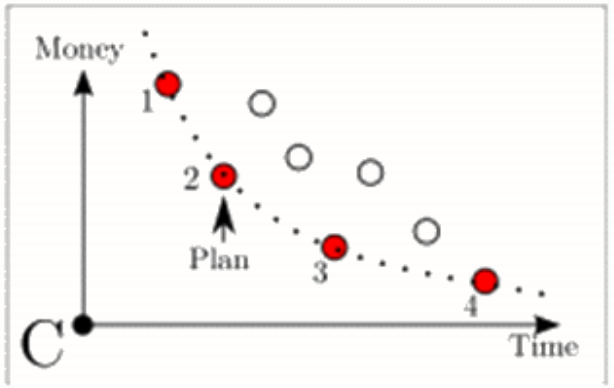


Problems and contributions

Selection of views to materialize



Multicriteria optimization



- Flexible pricing models
- Cost models for view materialization
- Detailed model of the optimization process

Transfer cost

$$C_t(D, Q, A) = C_t^-(D, Q) + C_t^+(A)$$

ascending descending

- *D* : Data set
- *Q* : Query workload
- *A* : Query result

Transfer cost

$$C_t(D, Q, A) = C_t^-(D, Q) + C_t^+(A)$$

ascending descending

$$\approx C_t^+(A)$$

- D : Data set
- Q : Query workload
- A : Query result

EC2

Volume	Cost
0 GB – 1 GB	0
1 GB – 10 TB	$0.12 / GB
10 TB – 40 TB	$0.09 / GB

Computation cost

$$C_c(Q, IC) = \sum_{i=1}^{n_Q} \sum_{j=1}^{n_{IC}} t(Q_i, IC_j) \times c_c(IC_j)$$

Processing time

Renting cost

- $Q = \{Q_i\} \mid i = 1..n_Q$: Query workload
- $IC = \{IC_j\} \mid j = 1..n_{IC}$: Configuration of computing instances

Storage cost

$$C_s(D) = \sum_{k=1}^{n_D} c_s\big(s(D_k)\big) \times t(D_k)$$

Storage cost

Storage duration

Data size

- $D = \{D_k\} \mid k = 1..n_D$: Stored data per periods of time

Storage cost

S3

Volume	Cost
0 TB – 1 TB	$0.140 / GB
1 TB – 450 TB	$0.125 / GB

$$C_s(D) = \sum_{k=1}^{n_D} c_s\left(s(D_k)\right) \times t(D_k)$$

Storage cost

Storage duration

Data size

- $D = \{D_k\} \mid k = 1..n_D$: Stored data per periods of time

Computation cost with materialized views

Processing time

$$C_c(Q,V,IC) = T(Q,V) \times c_c(IC_0) \times n_{IC}$$

Renting cost

- *Q* : Query workload
- *V* : Set of materialized views
- *IC* : Configuration of computing instances

Computation cost with materialized views

Processing time

$$C_c(Q,V,IC) = T(Q,V) \times c_c(IC_0) \times n_{IC}$$

Renting cost

$$T(Q,V) = T_{proc}(Q,V) + T_{mat}(V) + T_{maint}(V)$$

Query execution Materialization Maintenance

- *Q* : Query workload
- *V* : Set of materialized views
- *IC* : Configuration of computing instances

Storage cost with materialized views

Storage cost

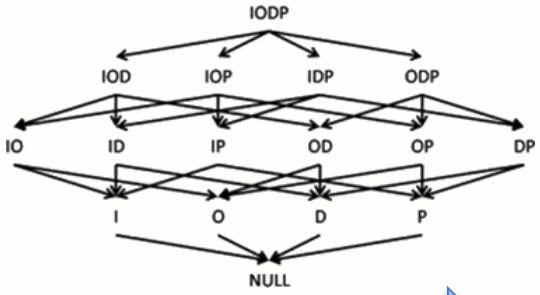
Storage duration

$$C_s(D,V) = c_s(s(D) + s(V)) \times t$$

Data size

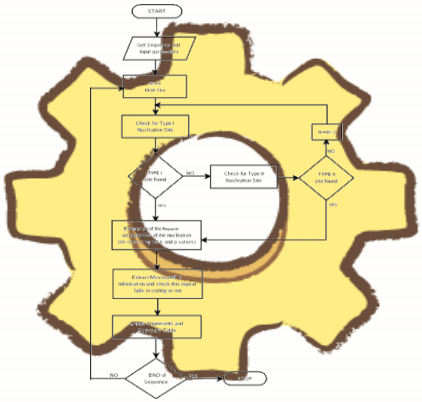
- D : Data set
- V : Set of materialized views

Optimization process



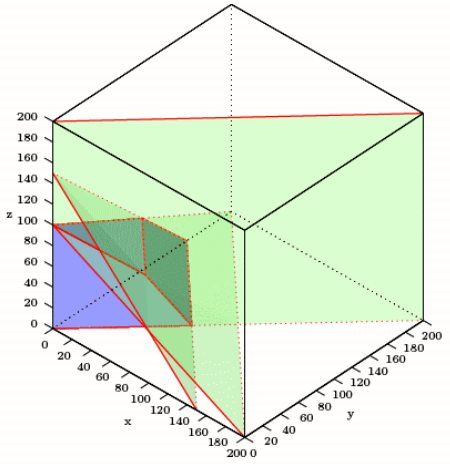
V_{all}

View selection
(existing algorithm)



V_{cand}

Linear program



V

Optimization problems

Find a set of materialized views $V \subseteq V_{cand}$

- *MV1*
  - Minimize T_{proc}
  - Constraint: $C \leq C_{max}$
- *MV2*
  - Minimize C
  - Constraint: $T_{proc} \leq T_{max}$
- *MV3*
  - Minimize $\alpha \times T_{proc} + (1 - \alpha) \times C$

Experimental environment


VM1


VM2


VM3

...


VM20

- 2 GB RAM
- 8 GB HDD
- Hadoop 0.20.2
- Pig 0.9.1


P1

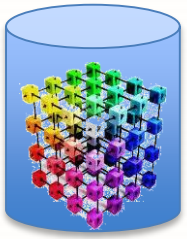

P2

...


P12

- Quadri-pros 800 MGhz
- 96 GB RAM

Experimental environment



Star Schema Benchmark

- Data: 5.5 GB
- 4 series of queries

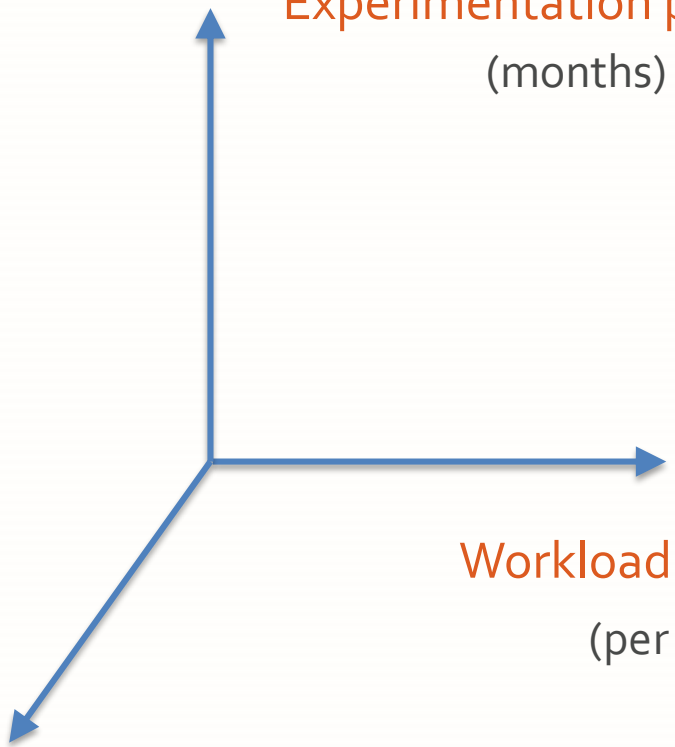
...

- 2 GB RAM
- 8 GB HDD
- Hadoop 0.20.2
- Pig 0.9.1

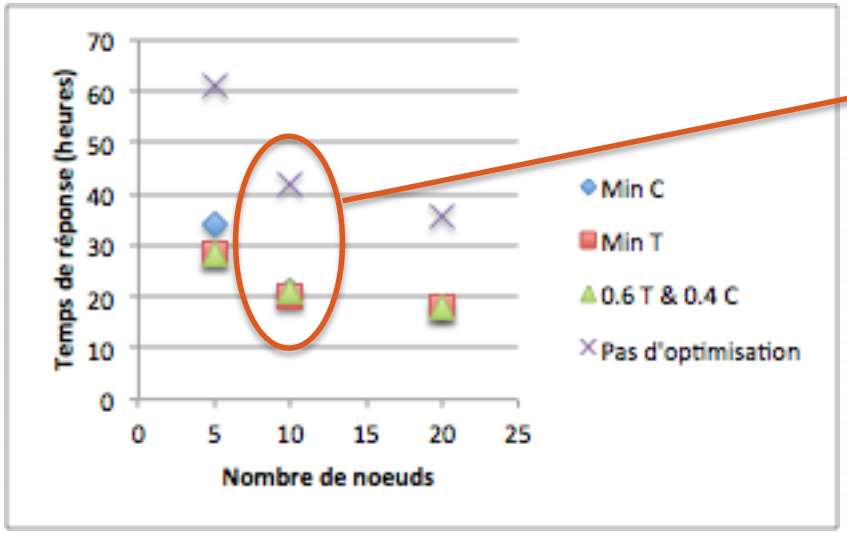
...

- Quadri-pros 800 MGhz
- 96 GB RAM

Parameters

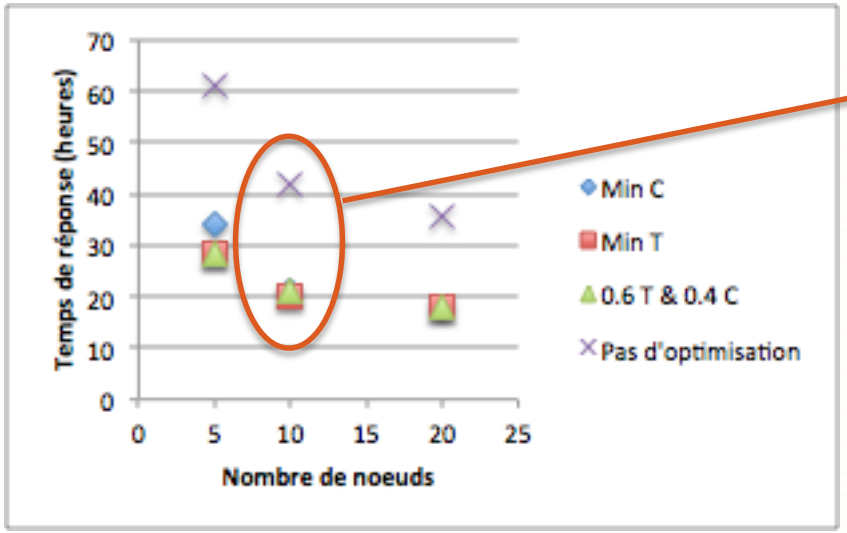

Experimentation period: 1-24 [12]
(months)
Workload frequency: 1-5 [4]
(per week)
Number of nodes: 5-20 [10]

Experimental results

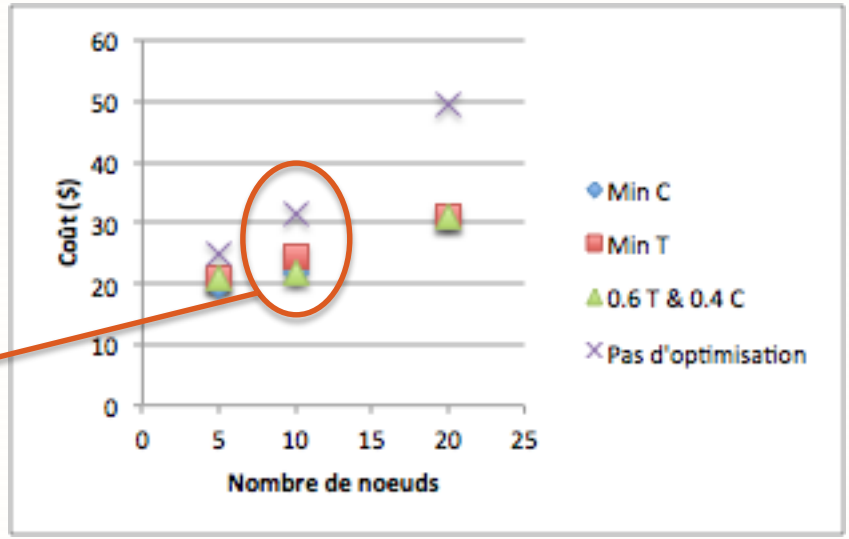


Performance gain: 110%

Experimental results



Performance gain: 110%



Cost gain: 30%

Results

- New cost models (cloud pricing)
- Multicriteria optimization process
- View materialization always desirable
- Optimization objectives not contradictory

Results and perspectives

- New cost models (cloud pricing)
- Multicriteria optimization process
- View materialization always desirable
- Optimization objectives not contradictory
- Enhance cost models
- Extend to other pricing models
- Integrate materialized view selection and optimization phases
- Exploit other optimization techniques
- Experiment on larger scales
- Better optimization algorithms

Results and perspectives

- New cost models (cloud pricing)
- Multicriteria optimizati… process
- View materialization desirable
- Optimization objectives not contradictory

- Enhance cost models
- Extend to other pricing models
- …egrate materialized view selection …optimization phases
- …it other optimization …niques
- Experiment on larger scales
- Better optimization algorithms