Cost Models for Materialized View Selection in the Cloud

Application to Amazon EC2 and S3 Services

Romain Perriot
Jérémy Pfeifer
Laurent d’Orazio
Bruno Bachelet
Sandro Bimonte
Jérôme Darmont
BI in the cloud(s)

SUNY Goodness

ITEC
(Library & Online Learning)

RF

OAS

Local Campus

Campus ERP

grades

enrollment

courses

curriculum

majors

demographics

programs

budget

Campus ERP

Information

Trends

Insight

Analysis

Projections

Knowledge

Decisions

EDA 2013 - Cost Models for Materialized Views in Amazon EC2 and S3
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
- ...

Introduction
Validation
Conclusion
Let’s be pragmatic!

Cloud service provider

\[
\text{Cost}_{\text{global}} = \text{Cost}_{\text{transfer}} + \text{Cost}_{\text{cpu}} + \text{Cost}_{\text{storage}}
\]

Indexes
Caches
Materialized views
Fragmentation
…

EDA 2013 - Cost Models for Materialized Views in Amazon EC2 and S3
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

\[ C_t(D, Q, A) = C_t^-(D, Q) + C_t^+(A) \]

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

\[ C_t(D, Q, A) = C_t^-(D, Q) + C_t^+(A) \]

\[ \approx C_t^+(A) \]

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

**EC2 Pricing View Materialization**

<table>
<thead>
<tr>
<th>Volume</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 GB – 1 GB</td>
<td>0</td>
</tr>
<tr>
<td>1 GB – 10 TB</td>
<td>$0.12 / GB</td>
</tr>
<tr>
<td>10 TB – 40 TB</td>
<td>$0.09 / GB</td>
</tr>
</tbody>
</table>
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\} / i = 1..n_Q : Query workload

**IC** = \{IC_j\} / j = 1..n_{IC} : Configuration of computing instances
Storage cost

\[ C_S(D) = \sum_{k=1}^{n_D} c_s(s(D_k)) \times t(D_k) \]

\[ D = \{D_k\} / k = 1..n_D : \text{Stored data per periods of time} \]
### Storage cost

**S3**

<table>
<thead>
<tr>
<th>Volume</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 TB – 1 TB</td>
<td>$0.140 / GB</td>
</tr>
<tr>
<td>1 TB – 450 TB</td>
<td>$0.125 / GB</td>
</tr>
</tbody>
</table>

\[
C_S(D) = \sum_{k=1}^{n_D} c_S(s(D_k)) \times t(D_k)
\]

- \( D = \{D_k\} / k = 1..n_D \): 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

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

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

\[ C_S(D, V) = c_s(s(D) + s(V)) \times t \]

- \( D \): Data set
- \( V \): Set of materialized views

**Introduction**

**Validation**

**Conclusion**
Optimization process

View selection
(existing existing algorithm)

Linear program

EDAS 2013 - Cost Models for Materialized Views in Amazon EC2 and S3
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

P1
P2
P12

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

- Quadri-pros 800 M Ghz
- 96 GB RAM
Experimental environment

Star Schema Benchmark

- Data: 5.5 GB
- 4 series of queries

VM1, VM2, VM3, VM20

P1, P2, P12

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

Quadri-pros 800 MHz
- 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%

EDA 2013 - Cost Models for Materialized Views in Amazon EC2 and S3
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)
- Enhance cost models
- Multicriteria optimization process
- Extend to other pricing models
- View materialization desirable
- Integrate materialized view selection optimization phases
- Optimization objectives not contradictory
- Experiment on larger scales
- Better optimization algorithms

EDA 2013 - Cost Models for Materialized Views in Amazon EC2 and S3