Approximate Answers to OLAP Queries on Streaming Data Warehouses

Michel DE ROUGEMONT, Phuong Thao CAO
Paris II Univ., South Paris Univ.
November 2, 2012

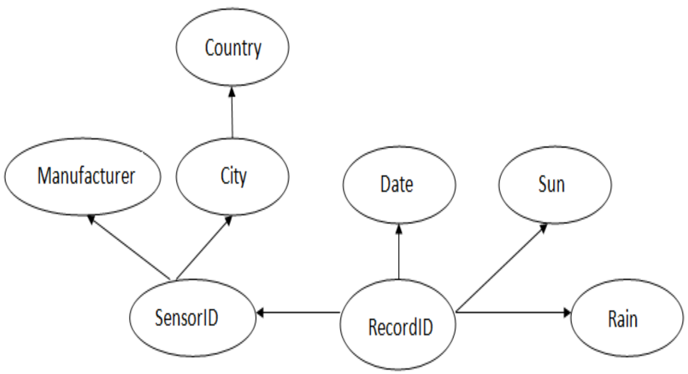
Outline

1. **Context: OLAP Queries**
   - Approximate answers
   - Streaming data
2. **Data exchange**

   Approximate answers with:
   - Sampling algorithm on the Sources
   - Use of Statistical dependencies
3. **Implementation**

1. Context

- OLAP Schema



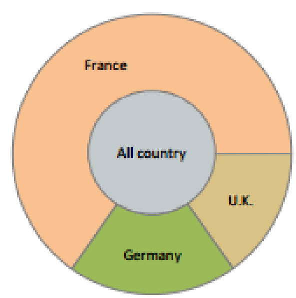
RecordID	SensorID	Date	Sun (hours)	Rain (hours)
1001	8	11/02/12	8	2
…	…	…	…	…

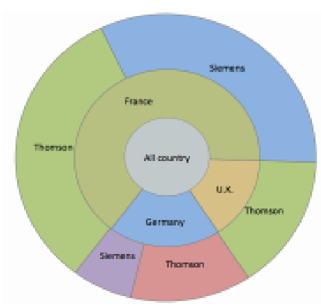
Fact table

- Different streams feed the Fact table

- OLAP queries

  (Sum of Measure)





Measure=Hours of Sun

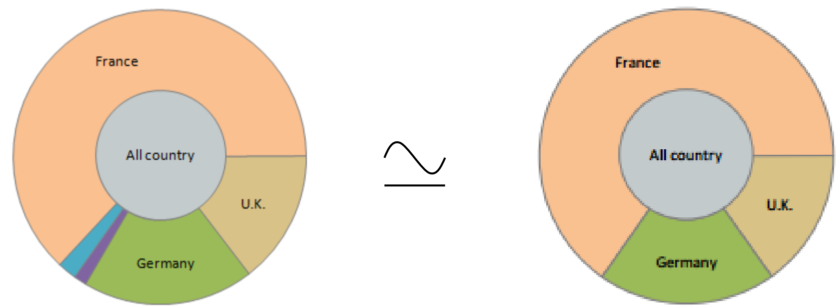
Analysis by Country

Analysis by Country/Manuf.

Approximation

- Distance L_1

  100% error for the blue area

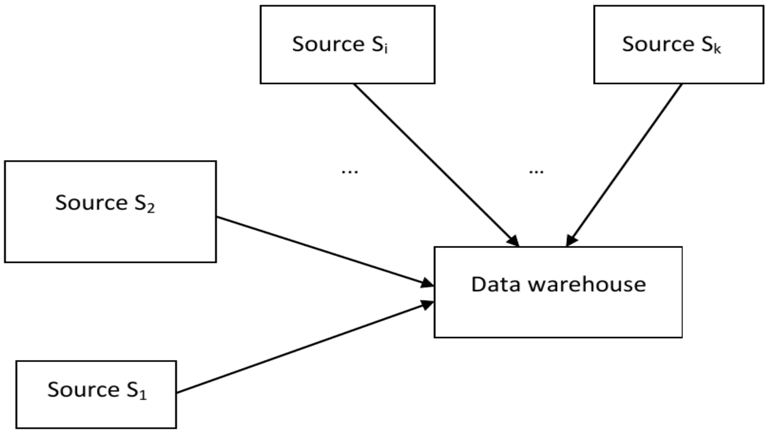


- Sampling:
  - classical technology to approximate
  - streaming: It is hard to approximate (Cormode et al. 2003)

- **Data Exchange**

2. OLAP Data Exchange

- Different sources



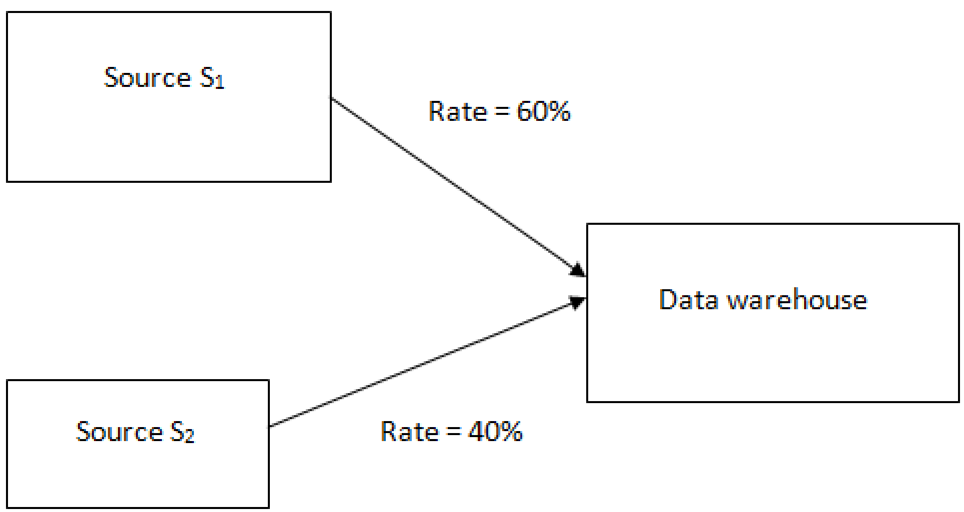
- Different streams: hard to approximate in the worst case (Cormode et al. 2012)

  How can we approximate queries in some special case? **statistical dependencies**

2.1 Streams with Different Rates

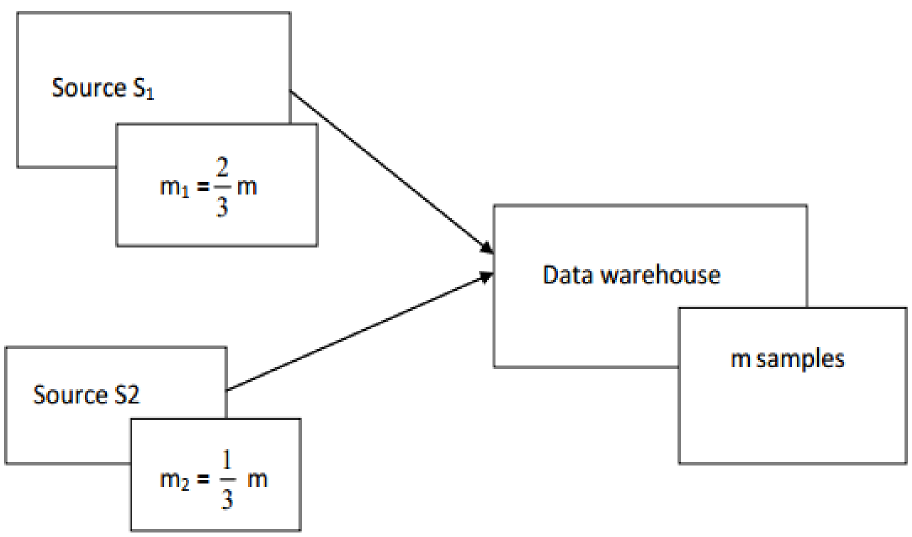
- Data warehouse
  - Union of different Sources
  - Rate of tuples of each Source is different

(rate: relative number of tuples per unit of times)



Uniform samples on the Streams

- Approximate Algorithm
  - Step 1: sampling on each Source with uniform distribution. **#samples % to the rate of the Source.**
  - Step 2: combine all samples according the rates
  - Step 3: approximation on the union of samples



Theorem: On a window of size T, OLAP queries are ε-approximated with N samples (which depend on T and ε) with high probability.

2.2 Special Case: Statistical Dependencies

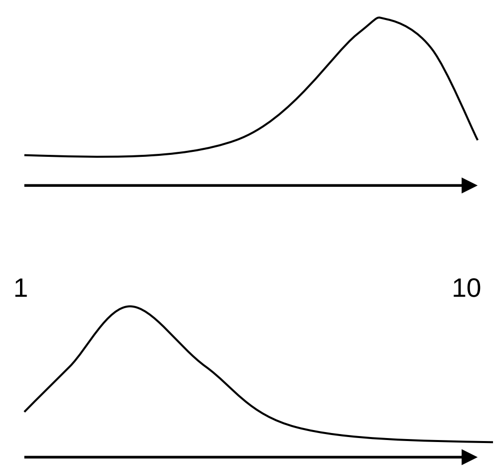
- Some attributes imply a distribution μ on the measure : A.B.C $\triangleleft$ M
  - (a,b,c) determines a fixed distribution on M
  - Generalization of functional dependencies

- City $\triangleleft$ Sun (μ distribution)

Hours of Sun

Marseille :

London :



Distribution of pairs

- City.Country

City	Country	Density of tuples
London	U.K	1/12
Berlin	Germany	1/12
Paris	France	1/6
…	…	…

→

Country	Distribution of Sun
U.K.	0.64
Germany	0.21
France	0.15

- Manufacturer.City (δ)

Manufacturer	City	Density of tuples
Thomson	London	1/12
Thomson	Berlin	1/12
Siemens	Paris	1/12
…	…	…

→

Manufacturer	Distribution of Sun
Siemens	0.39
Thomson	0.61

Use of Statistical Hypothesis:

Distributed Algorithm

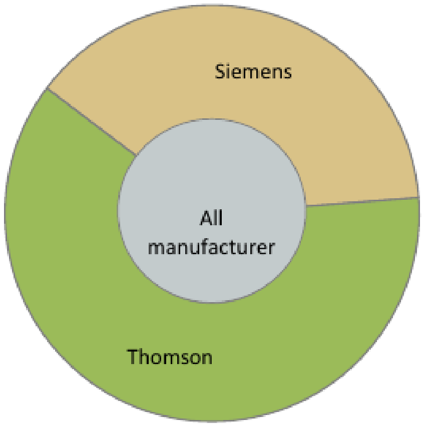
- **Each Source i, we sample by uniform distribution and:**
  - Learn the μ_i
  - Estimate the distribution on pairs δ_i
  - Estimate its rate: r_i
- **Data Warehouse:**
  - Combine rates r_i, δ_i and μ_i to approximate the OLAP query on A (Manufacturer)

$$Q^{M}_{C=Siemens} = (r_1.Q^{M}_{C=Siemens})^1 + (r_2.Q^{M}_{C=Siemens})^2$$

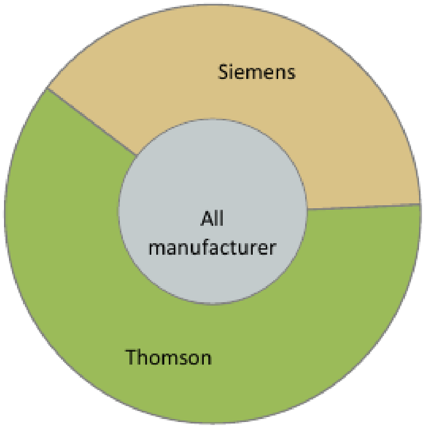
$$= \frac{2}{3}.\left[\sum_{City} \delta(Siemens, City).Avg(\mu_{City})\right] + \frac{1}{3}.\left[\sum_{City} \delta(Siemens, City).Avg(\mu_{City})\right]$$

$$= 0.39$$

Statistical Model



Exact answer



Approximate answer

Advantages:

- Statistical dependencies : more intuitive
- Sources send only statistical dependencies

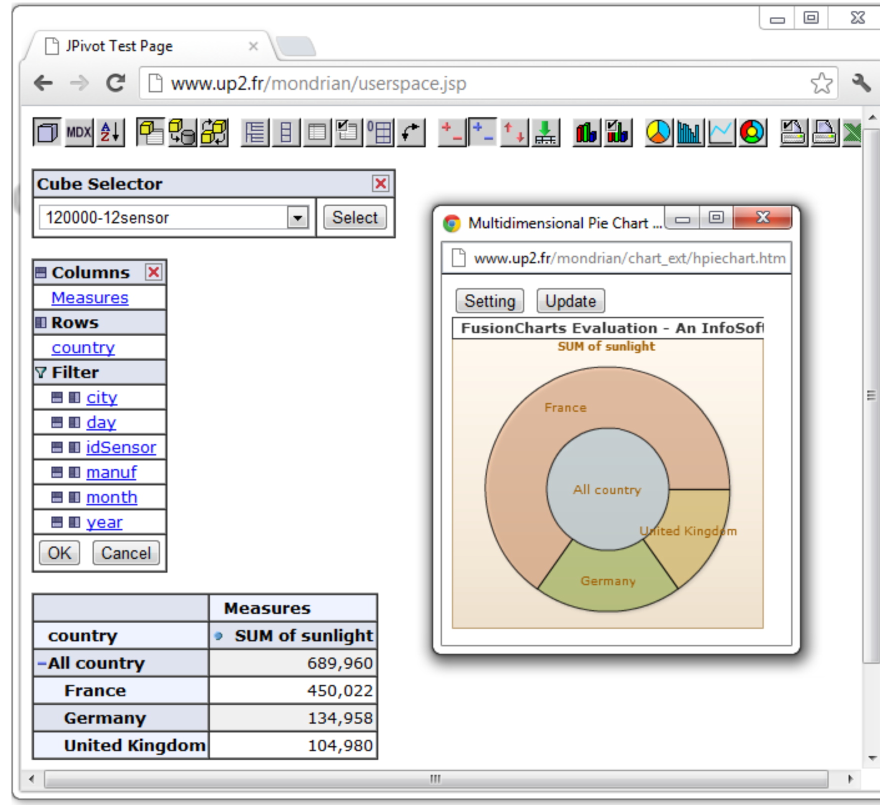
  (constant size of information on finite domains)
- Sources do not send samples

Our contribution

- Special situation: model of statistical dependencies on streaming data
- Approximation algorithms:
  - Sampling: each Source samples and we combine all the samples
  - Statistical model: combine statistical dependencies and distributions of pairs
- Worst case is not approximable

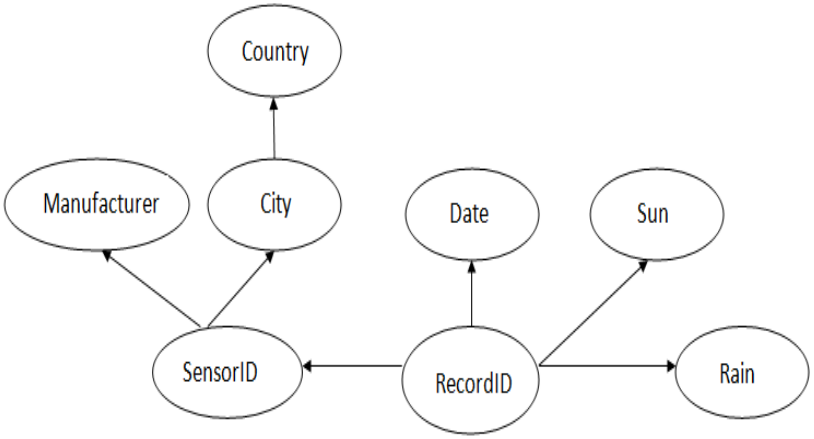
3. Implementation

- Program
  - Mondrian OLAP engine
  - Jpivot interface
- Data warehouse
  - 10^6 tuples

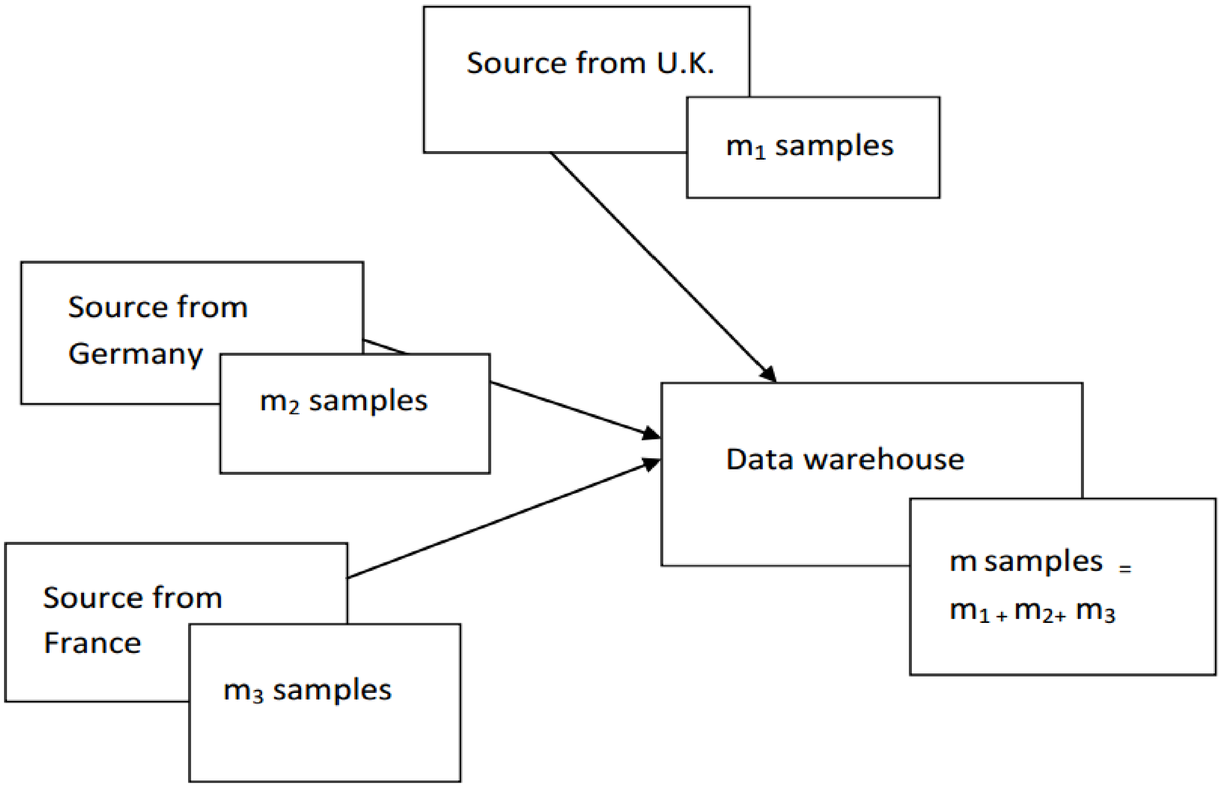


Approximate answer on sources:

- **Data warehouse**
  - 12 sensors: 6 in France, 3 in Germany, 3 U.K.
  - 2 manufacturers: Siemens, Thomson
  - 9 cities
  - 1≤ Sun, Rain ≤10
  - Statistical dependencies:
    - City $\lhd$ Sun
  - Distribution of pairs
    - City.Country
    - Manufacturer.City



Example 1: Analysis by country



Approximate answer on sources:
Analysis by country

- Learn **distributions μ_i , δ_i** from samples

City	Average value of Sun : $Avg_{\mu}(a_i)$
London	3.5
Berlin	5
Paris	7.5
…	…

City	Country	Density of tuples : $\delta(a_i)$
London	U.K.	1/12
Berlin	Germany	1/12
Paris	France	1/6
…	…	…

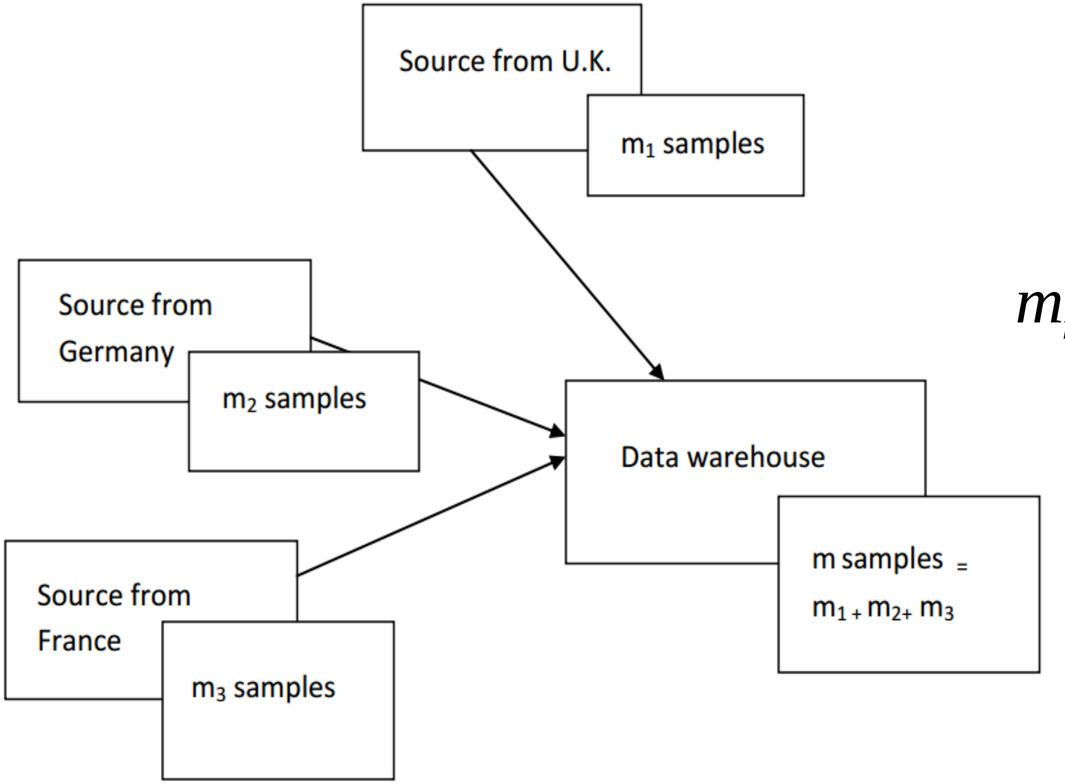
→

Country	Distribution of Sun
U.K.	0.64
Germany	0.21
France	0.15

$$m_i = m \times \frac{\delta(a_i) \times Avg_{\mu}(a_i)}{\sum_i \delta(a_i) \times Avg_{\mu}(a_i)}$$

Approximate answer on sources:

Analysis by country



$$m_i = m \times \frac{\delta(a_i) \times Avg_{\mu}(a_i)}{\sum_i \delta(a_i) \times Avg_{\mu}(a_i)}$$

Example 2: Analysis by Manufacturer

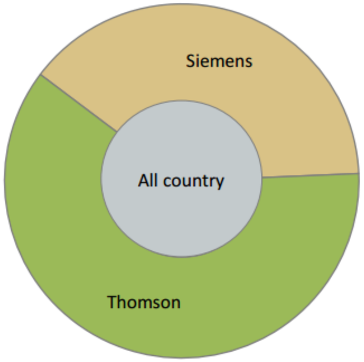
City	Avg value of Sunlight : Avg $_{\mu}(a_i)$
London	3.5
Berlin	5
Paris	7.5
…	…

Manufacturer	City	Density of tuples : $\delta(a_i)$
Thomson	London	1/12
Thomson	Berlin	1/12
Siemens	Paris	1/12
Thomson	Paris	1/12
…	…	…

→

Manufacturer	Distribution of Sun
Siemens	0.39
Thomson	0.61

↓



Approximate answer:
Analysis by Manuf.

Analysis of errors

	Distribution of Answers			
Manufacturer	Uniform sampling	Measure-based sampling	Linear estimation by the data exchange	Exact answer
Siemens	0.3851	0.4100	0.3890	0.3911
Thomson	0.6149	0.5900	0.6110	0.6089
TOTAL ERROR	0.0120	0.0378	0.0042	

- All algorithms: rate of errors < 4%
- Statistical model is better than uniform sampling
- Statistical model is better than Measure based sampling

Conclusion and Perspective

- **Conclusion**
  - In the case of statistical dependencies, the algorithm keeps a good approximation to OLAP queries
  - Constant information exchanged on finite domains
  - Required memory in the worst case: $\Omega(N)$
- **Perspective:**
  - Application to RSS
  - Decision tree for the statistical model: discover the statistical dependencies

Thank you !

Questions & Answers?