
Query Processing and Optimization on the Web

Mourad Ouzzani and Athman
Bouguettaya

Presented By Issam Al-Azzoni

Outline

- Part 1
 - Introduction – Web Data Integration Systems
 - Query Optimization in Web Data Integration Systems
 - Classification of Query Optimization Techniques in Web Data Integration Systems

Outline

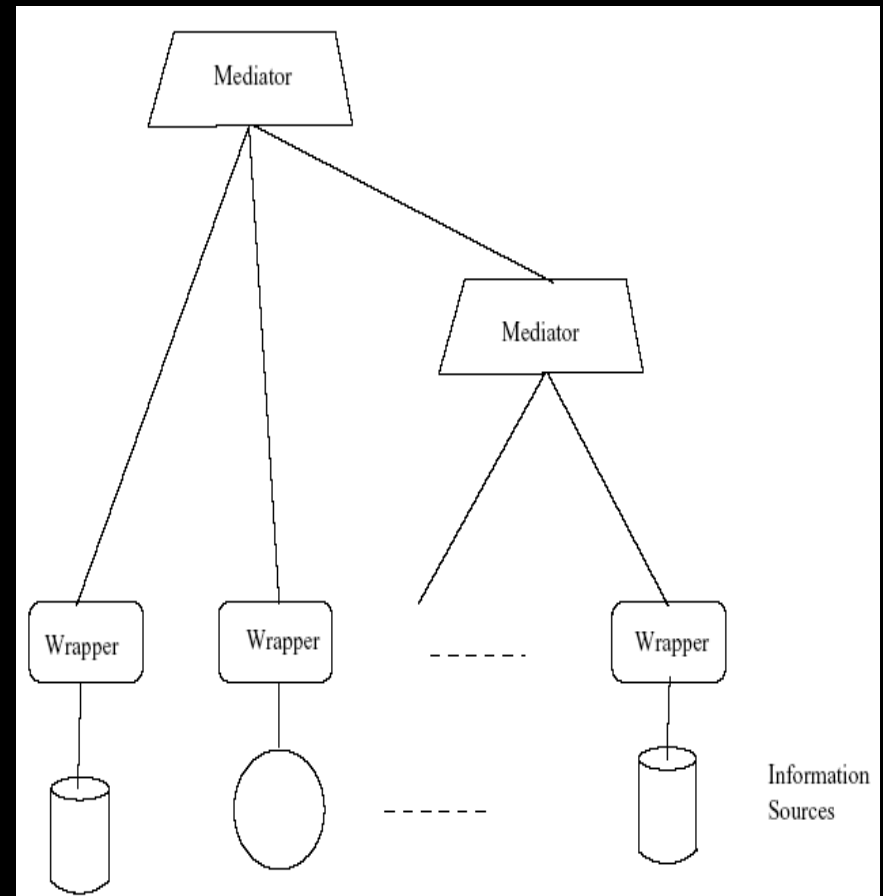
- Part 2
 - Garlic
 - ObjectGlobe
 - Interleaving Query Scheduling and Optimization
 - Tsimmis
- Part 3
 - Discussion

Introduction

- Integrating information from the Web
 - Large number of data sources
 - The dynamic nature of data sources
 - Heterogeneity of data sources
- There is a need to provide support to access information of the Web uniformly (**view integration**) and efficiently (**query optimization**)

View Integration

- The Mediator/Wrapper architecture
 - A wrapper for each data source
 - The mediator
 - Data source selection
 - Query translation
 - Result merging
 - Unified view of data



Query Optimization

- Very important for Web data integration systems
- Challenges:
 - The nature of the Web (heterogeneity, autonomy, large user base, ...)
 - The optimization objective functions (response time, quality of information, network sources, ...)
 - Locating appropriate data sources

Dimensions for Query Optimization on the Web

- Scalability
- Autonomy preservation
- Optimization performance
- Adaptiveness
- Capabilities

Query Optimization Techniques in Web Data Integration Systems

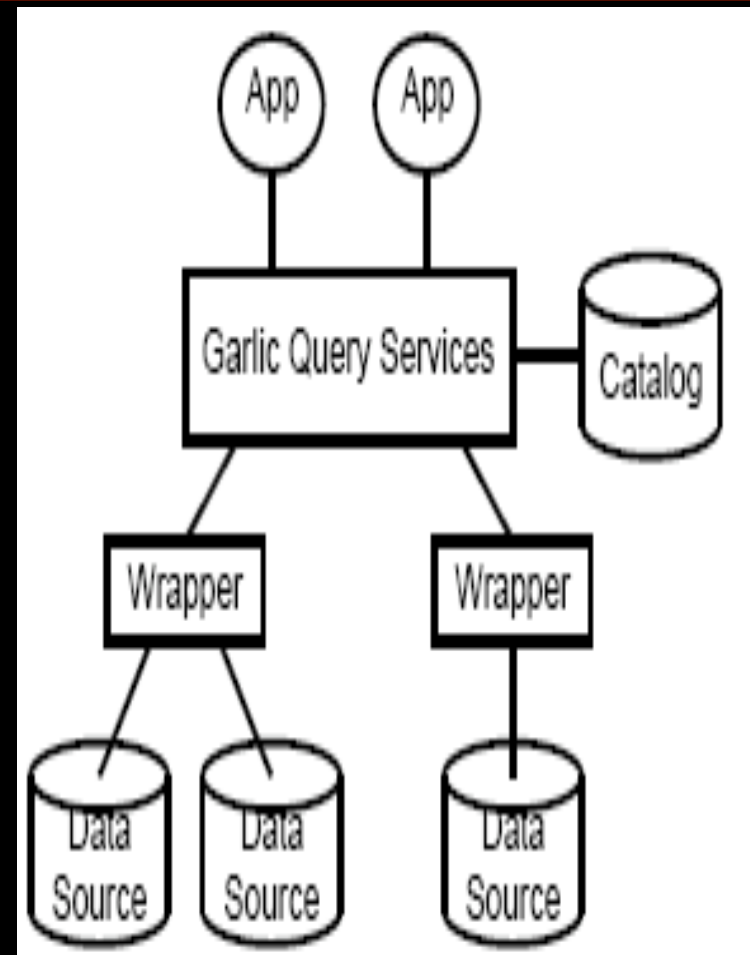
- Cost-based (Disco, Garlic, ...)
- Quality-based (ObjectGlobe, HiQIQ, ...)
- Adaptive query optimization (Telegraph, Tukwila, ...)
- Capability-based (Tsimmis, Infomaster, ...)

Cost-based Query Optimization Techniques

- Assume the existence of the necessary statistics
- Wrappers export statistics
- Query processor and wrappers cooperate to estimate costs
- Query processor gathers statistics advertised by sources
- Cache statistics

Garlic

- Data model: GDL (a variant of ODMG Object Description Language)
- Plans in Garlic are trees of operators, or POPs (Plan OPerators)

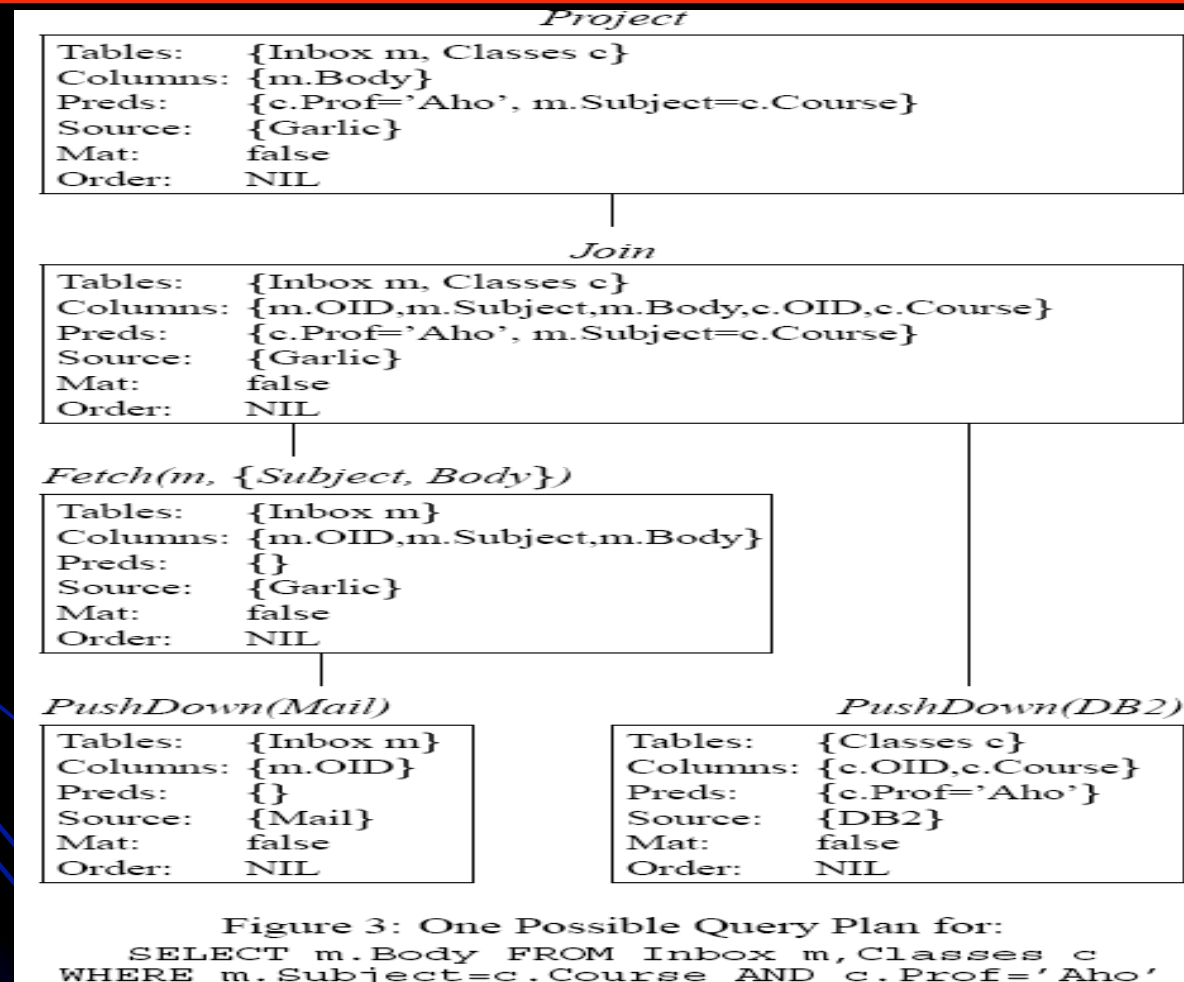


Garlic

- Plans are characterized by a set of plan properties

Property	Description
<i>Tables</i>	set of tables that have been accessed and joined
<i>Columns</i>	set of columns of the output of the plan
<i>Preds</i>	set of predicates that have been applied in the plan
<i>Source</i>	where the output is produced; i.e., the <i>id</i> of a data source or Garlic's execution engine
<i>Mat</i>	TRUE if the output of the plan is materialized; FALSE otherwise
<i>Order</i>	a sort expr. if the tuples of the output are ordered; NIL otherwise
<i>Cost</i>	estimated cost of the plan
<i>Card</i>	estimated number of tuples of the output of the plan

Garlic



Garlic

- Wrappers cooperate in the estimation of the total cost of a given query plan
- The cost of a plan is the sum of local processing costs, communication costs, and the costs to initiate subqueries

Quality-based Query Optimization Techniques

- Use of quality information
- User Quality Constrains

ObjectGlobe

- Underlying information sources: Structured and semistructured data
- Sources description: RDF
- Data model: XML based model
- Query language: Extension of SQL
- Query Optimization: Query processing capabilities + QoS based optimization

ObjectGlobe

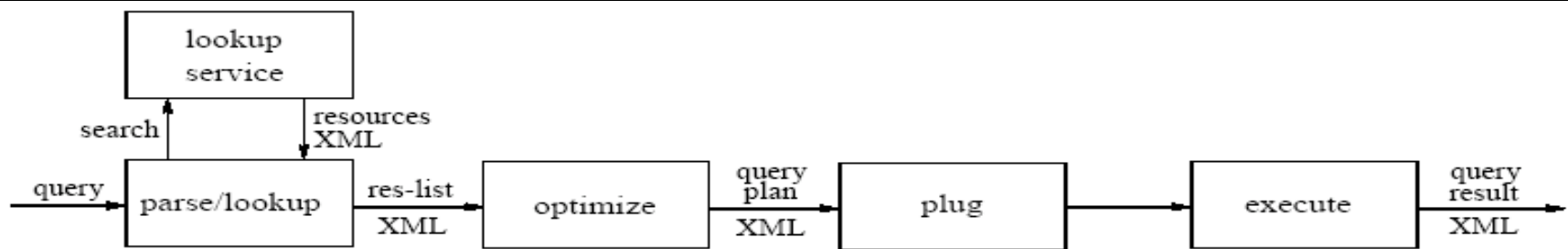


Fig. 1. Processing a query in ObjectGlobe

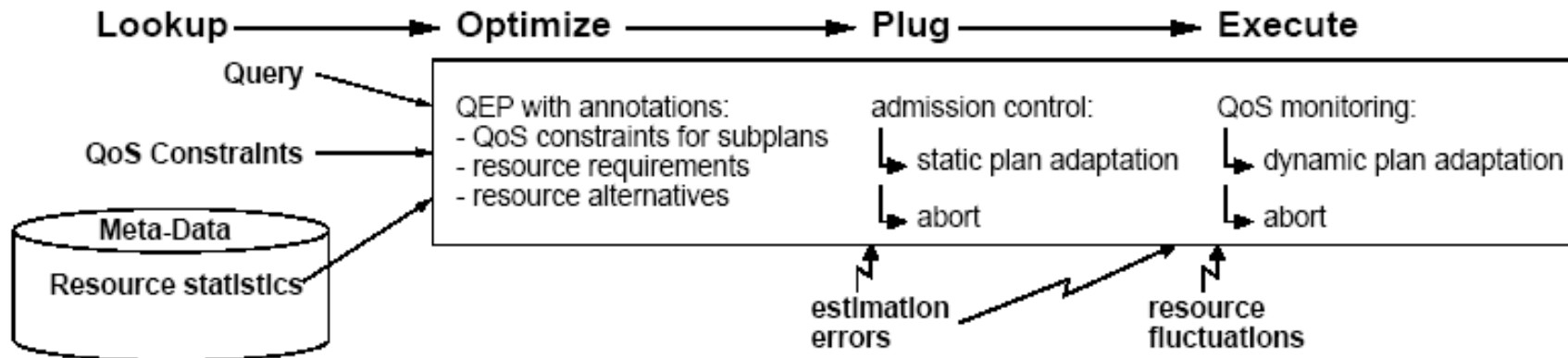


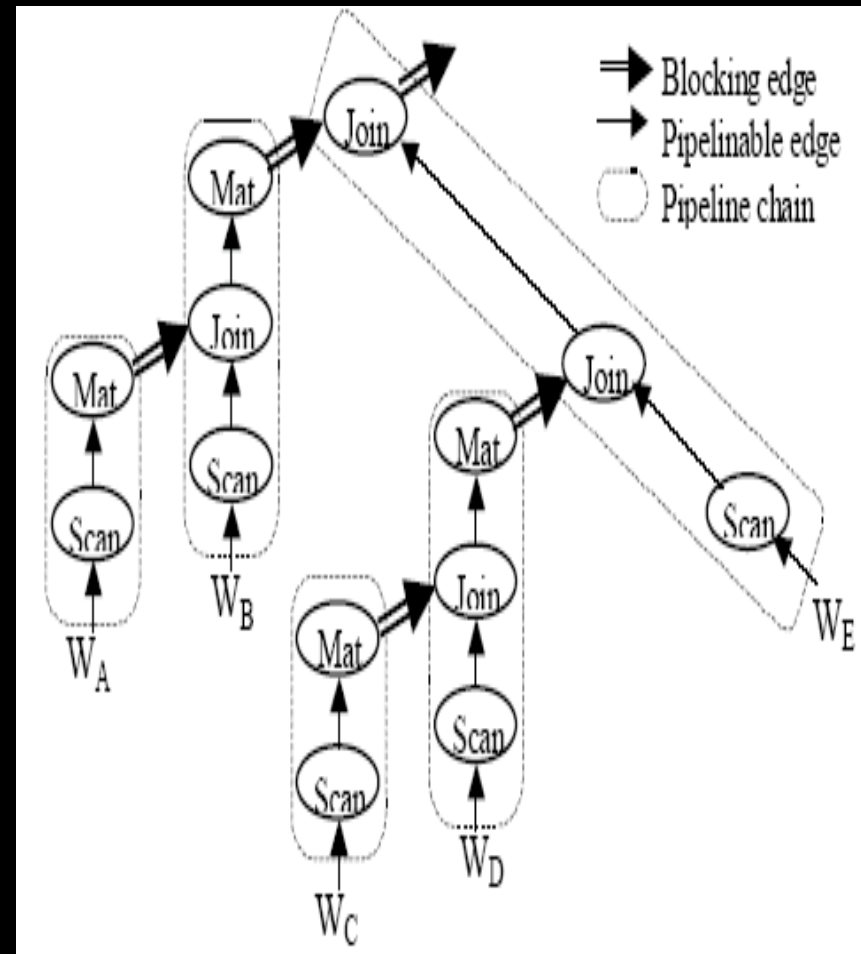
Fig. 6. The interaction of query processing and QoS management

Adaptive Query Optimization

- Data-flow recording
- Re-scheduling
- Re-optimization
- Use of special operators
- A query processor is adaptive if:
 - It receives information from its environment
 - It uses this information to determine its behaviour
 - This process iterates over time, generating a feedback loop between environment and behaviour

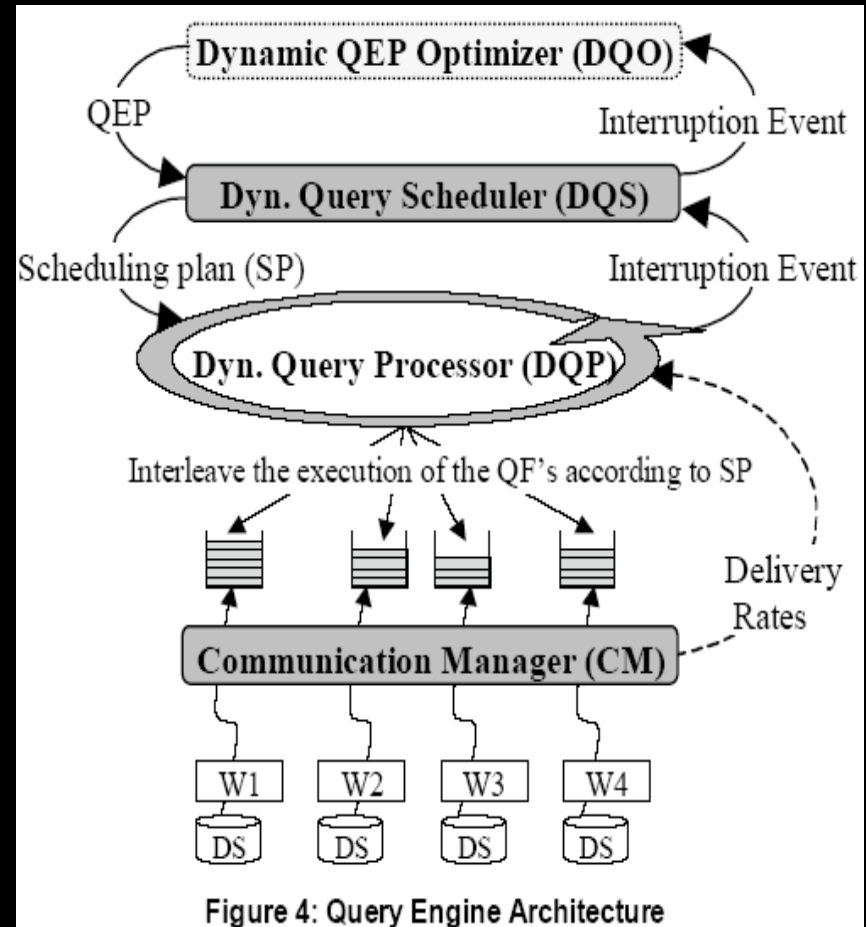
Interleaving Query Scheduling in Data Integration Systems

- Monitor the arrival rates at the data sources
- In the cases of significant changes, the execution plan is revised
- The query execution plan is represented by an operator tree with two kinds of edges:
 - A blocking edge
 - A pipelined edge



Interleaving Query Scheduling in Data Integration Systems

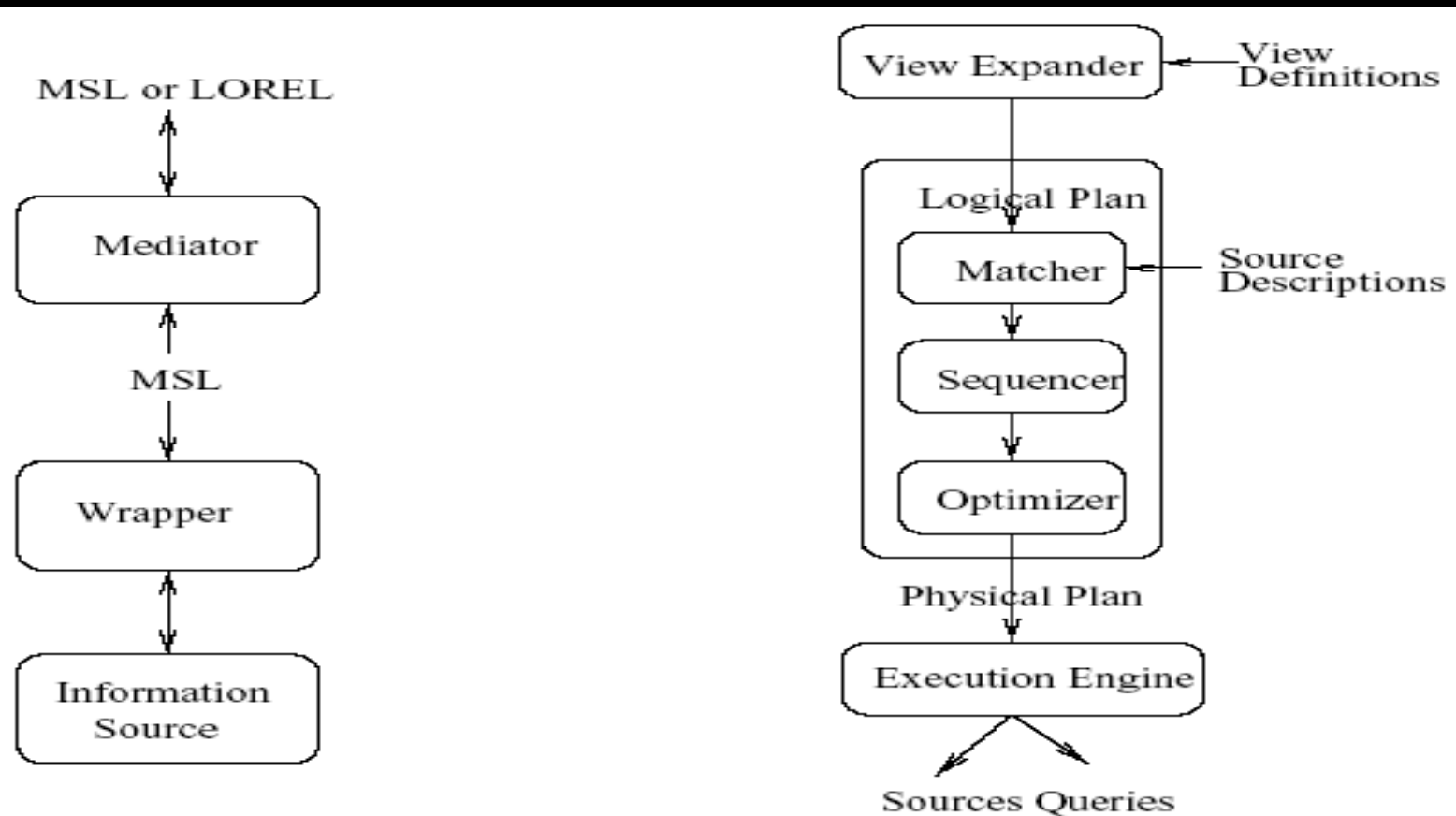
- Goal: Reduce the impact of unpredictable data delivery rate problems by scheduling several query fragments concurrently



Optimizing Queries over Sources with Limited Capabilities

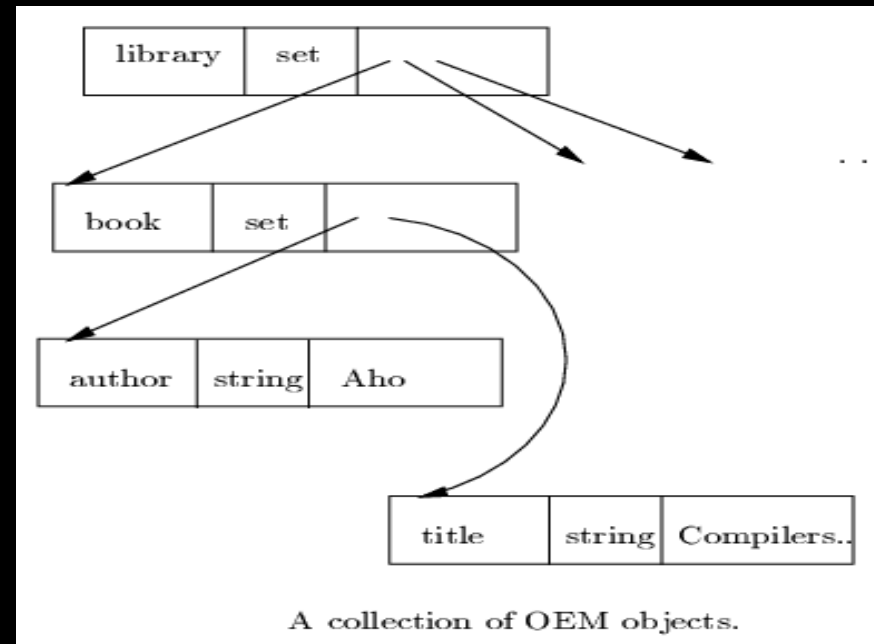
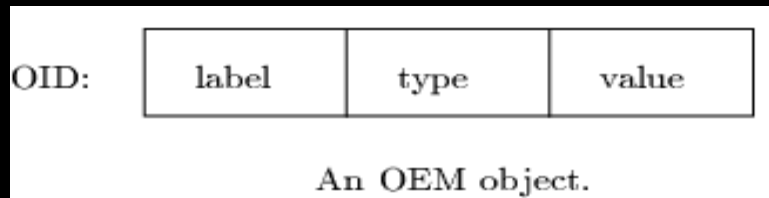
- How to deal with the often limited query capabilities of data sources:
 - Performance
 - Security
 - Legacy
 - Business

Tsimmis



Tsimmis architecture and query processor.

Tsimmis



Example : “Find the books of which Aho is an author.”

`<booktitle X> :-`

`<library { <book {<title X> <author "Aho">}}> }>@s1`

Future Trends

- The Semantic Web
 - Make information on the Web more “machine friendly”
- Web Services
 - A piece of software that can be defined, described and discovered by XML artifacts
- Issues:
 - Web service-based query language
 - Computational model
 - Optimization model

Discussion

- The presented classification seems to apply to Web data integration systems, rather than the optimization techniques
 - ObjectGlobe employs different query optimization techniques
- Should a Web data integration systems use all of the techniques?
 - How is this going to affect the optimizer?

Discussion

- There is no comparison between the different alternatives:
 - The query optimization techniques
 - The data models
 - The query languages
 - The mediation approaches
 - etc

Discussion

- Some concepts are not clearly explained
 - Source-centric vs query-centric mediation
- Can we learn from our knowledge in Heterogeneous Distributed Database Systems?
- Why is there a paradigm shift towards Web services?

References

- [1] L.M. Haas, D. Kossmann, E.L. Wimmers, and J. Yang, “Optimizing queries across diverse data sources,” in Proceedings of the 23rd International Conference on Very Large Data Bases (VLDB), Athens, Greece, Aug. 1997.
- [2] R. Braumandl, M. Keidl, A. Kemper, D. Kossmann, A. Kreutz, S. Seltzsam, and K. Stocker, “ObjectGlobe: Ubiquitous query processing on the internet,” The VLDB Journal, vol. 10, no. 1, 2001.
- [3] L. Bouganim, F. Fabret, C. Mohan, and P. Valduriez, “Dynamic query scheduling in data integration systems,” in Proceedings of the 16th International Conference on Data Engineering, San Diego, CA, USA, Feb.–March 2000.
- [4] M.T. Ozsu and P. Valduriez, Principles of Distributed Database Systems, Prentice Hall, 1999.

References

- [5] H. Garcia-Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, Y. Sagiv, J.D. Ullman, V. Vassalos, and J. Widom, “The TSIMMIS approach to mediation: Data models and languages,” *Journal of Intelligent Information Systems*, vol. 8, no. 2, 1997.
- [6] S. Thacker, A. Sheth, and S. Patel. “Complex Relationships for the Semantic Web.” <http://citeseer.ist.psu.edu/thacker01complex.html>
- [7] Aaron Sqartz, “The Semantic Web in Breadth.” <http://logicerror.com/semantic-web-long>

Class Discussion
