Architectural Solutions in Data Integration

Mustafa Jarrar

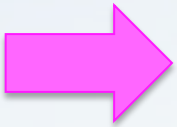
Birzeit University

Watch this lecture and download the slides

Online Courses : http://www.jarrar.info/courses

Thanks to Anton Deik for helping me preparing this lecture

Architectural Solutions in Data Integration



Part 1: Application-driven Integration Architectures

Part 2: Information Integration Architectures

Part 3: What Integration Criteria to Use

Keywords: Data Integration, Application-driven Integration, Data-driven Integration, Web Services, RPC, Publish & Subscribe, Consolidation, Data Warehouse, Data Integration, Service Oriented Architecture , Virtual Data Integration, Query complexity, heterogeneity

Different Solutions

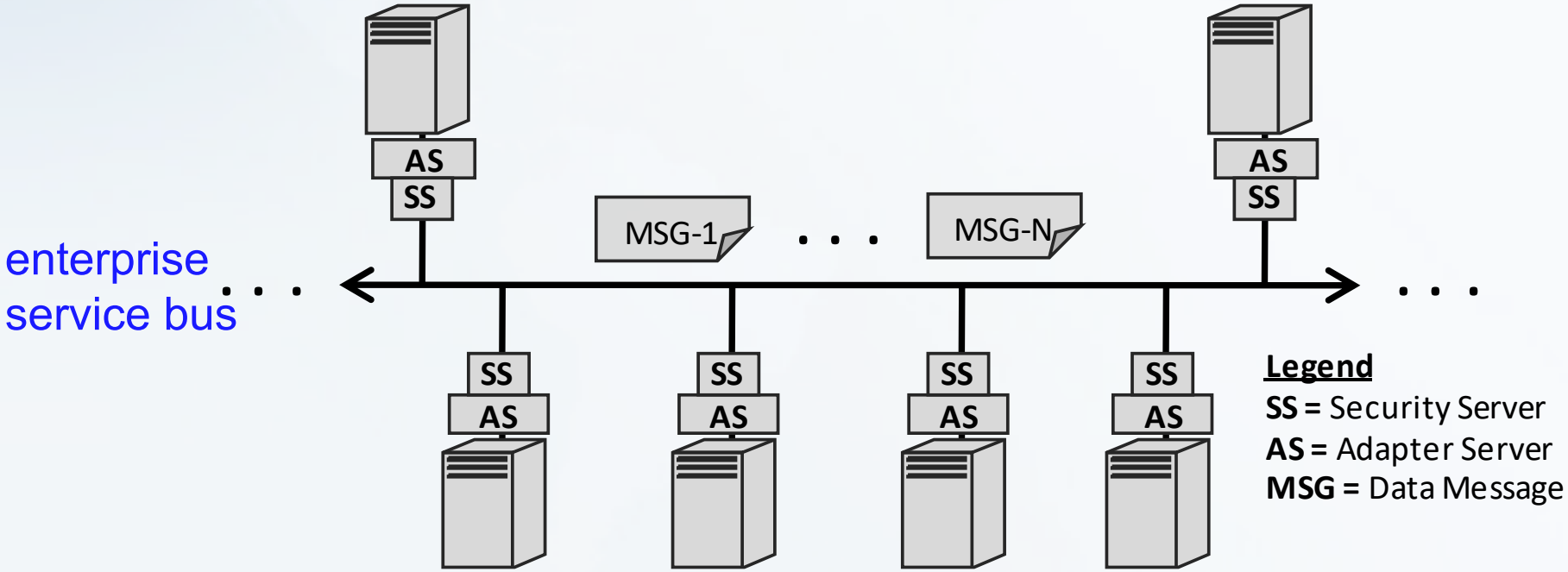
Two families of solutions for the integration issue:

- **Application-driven Integration**
  - Various types of middleware (e.g. Web Services, Remote Procedure Call (RPC), Publish & Subscribe) that achieve reconciliation through application to middleware communication

- **Data-driven Integration**
  - Various types of data reconciliation and integration
    - Consolidation
    - Data Fusion
    - Data Integration

Application-driven Integration

1- Service Oriented Architecture Scenario

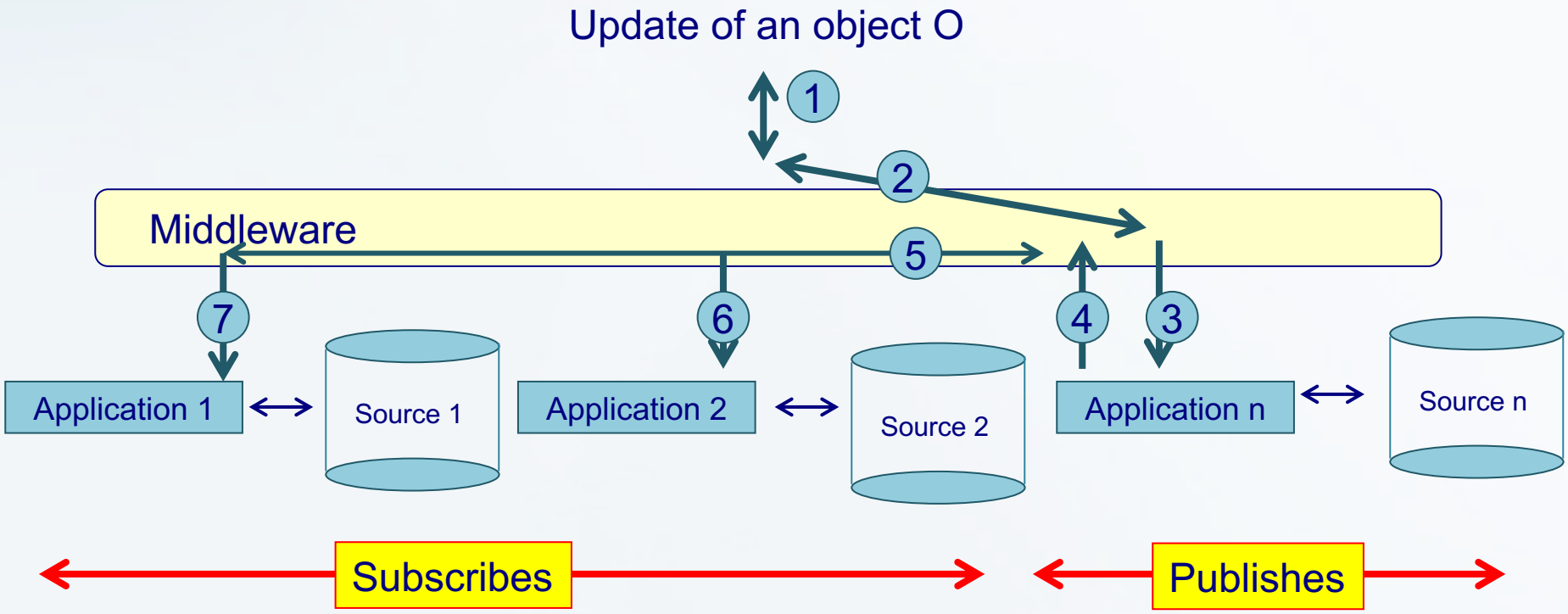


Application-driven Integration

Based on Carlo Batini [13]

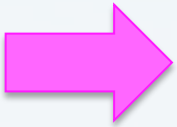
2- Publish-Subscribe Architecture Scenario

- Update via the middleware, then publish this update, other application that subscribe to receive updates, will also update their sources.
- Typical application-driven integration architecture for integration of updates.



Architectural Solutions in Data Integration

Part 1: Application-driven Integration Architectures



Part 2: Information Integration Architectures

Part 3: What Integration Criteria to Use

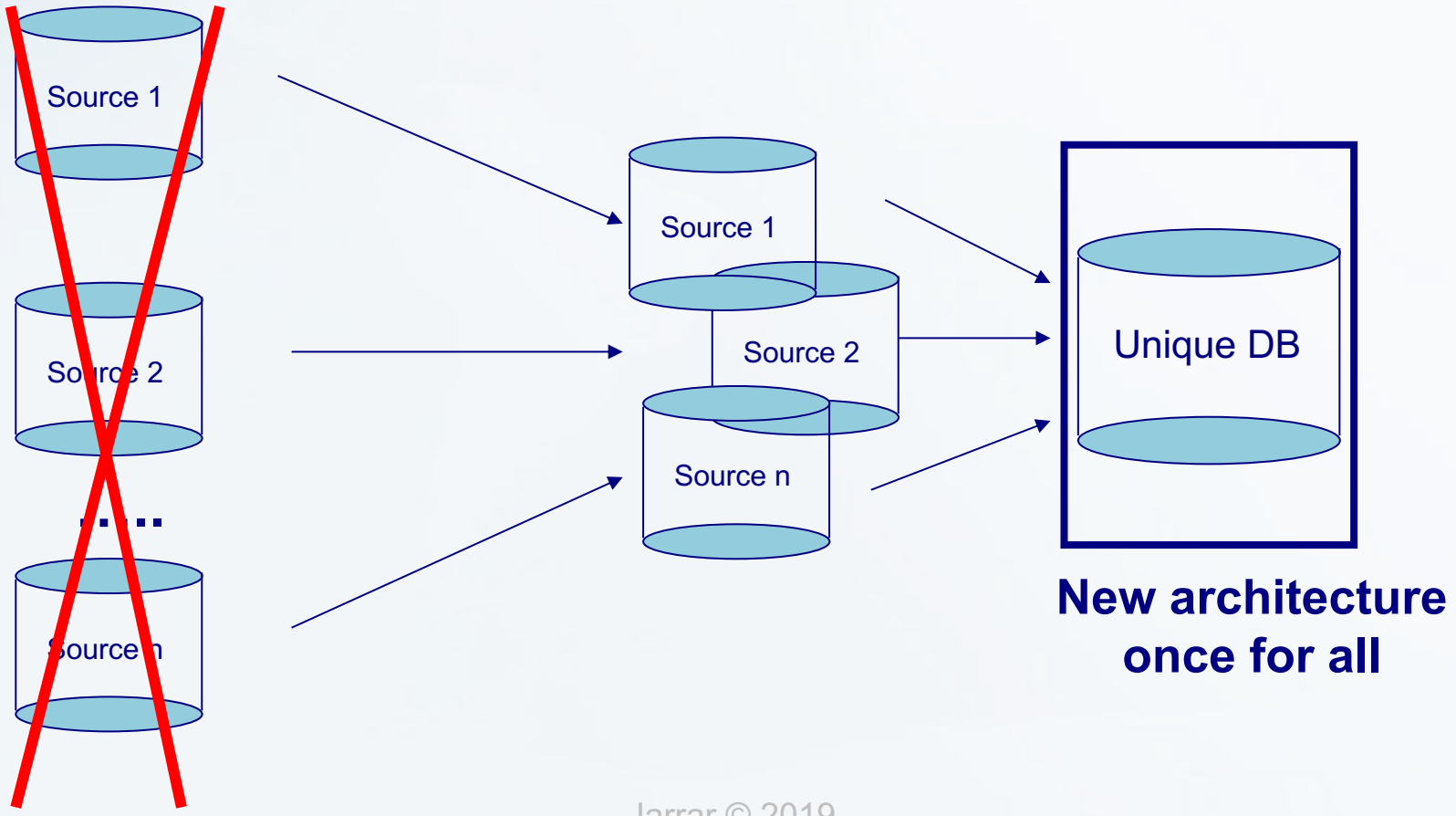
Keywords: Data Integration, Application-driven Integration, Data-driven Integration, Web Services, RPC, Publish & Subscribe, Consolidation, Data Warehouse, Data Integration, Service Oriented Architecture , Virtual Data Integration, Query complexity, heterogeneity

Information Integration Architectures

Based on Carlo Batini [13]

3- Consolidation Scenario

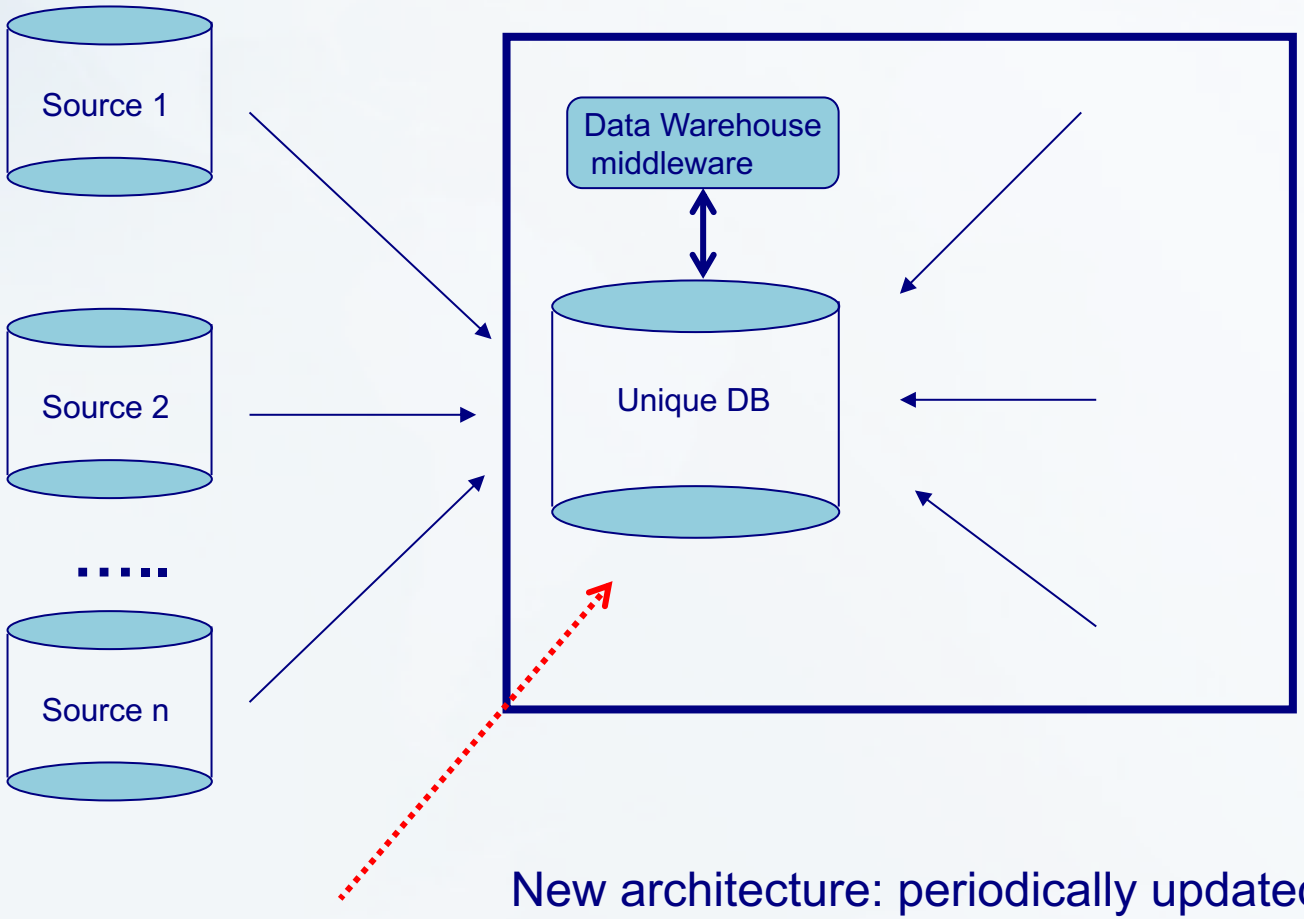
Merage all data sources into one new schema, and drop the old



Information Integration Architectures

Based on Carlo Batini [13]

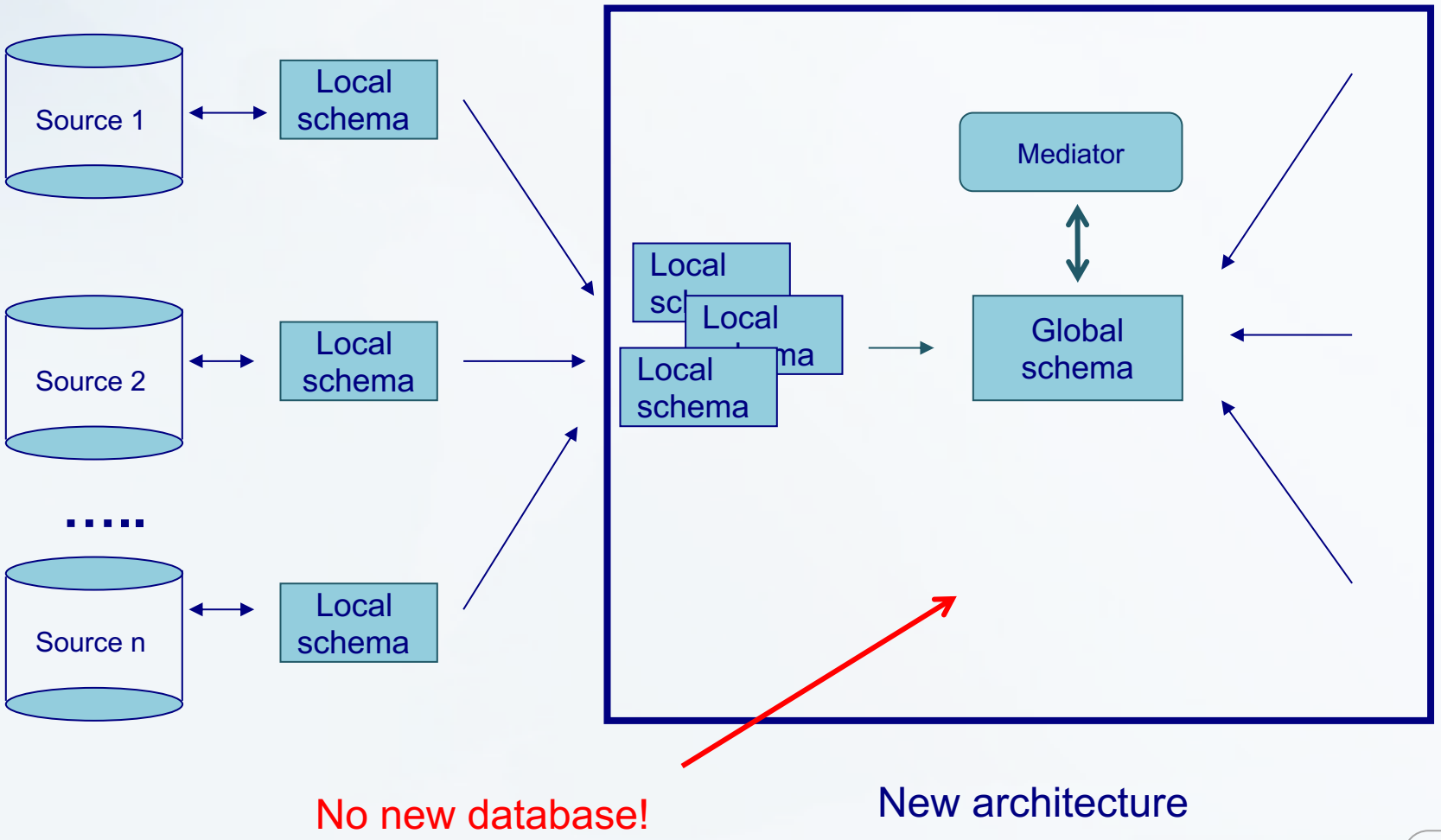
4- Data Warehouse Scenario



Information Integration Architectures

Based on Carlo Batini [13]

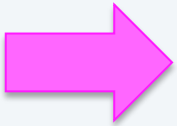
5- Virtual Data Integration Scenario



Architectural Solutions in Data Integration

Part 1: Application-driven Integration Architectures

Part 2: Information Integration Architectures

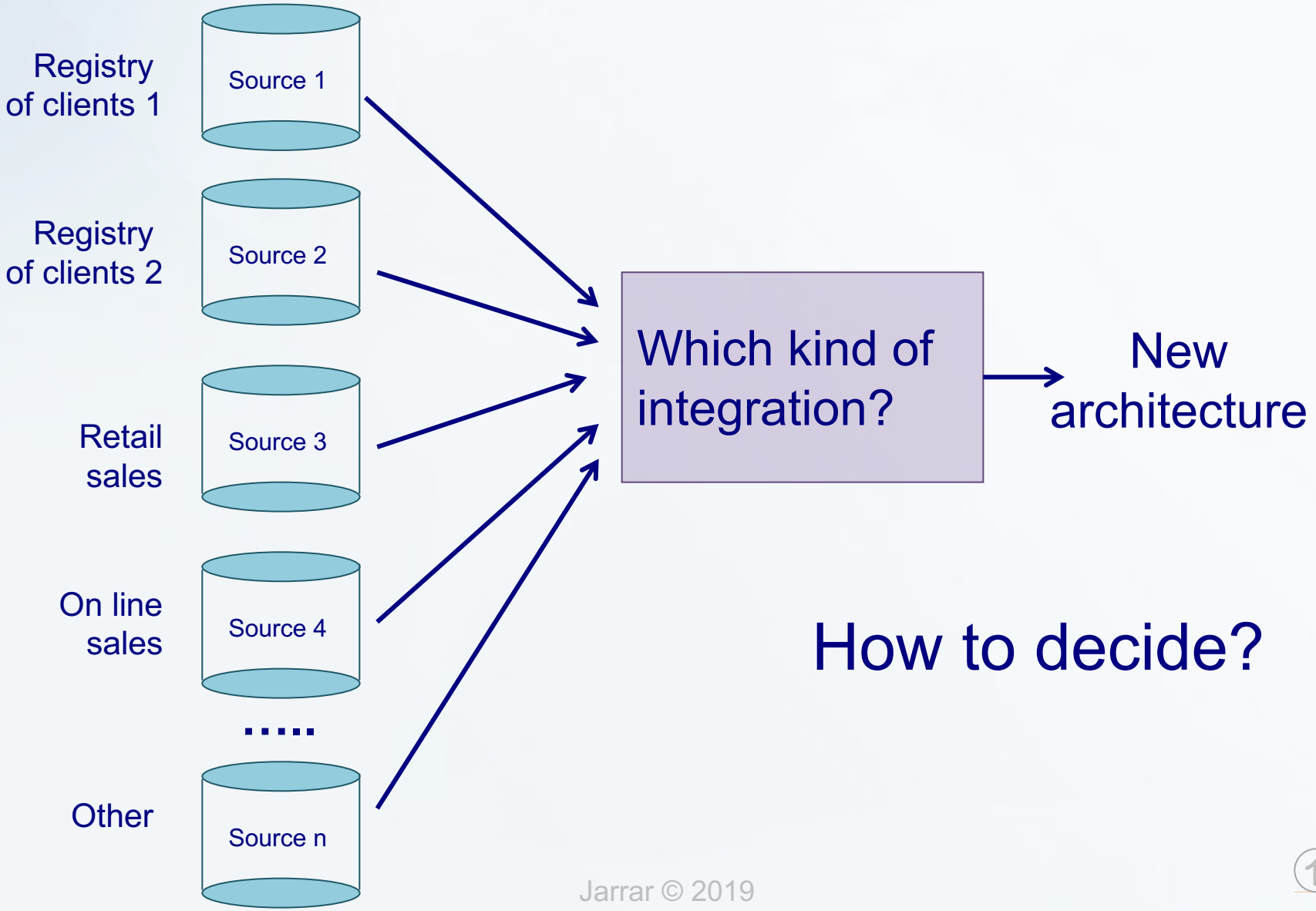


Part 3: What Integration Criteria to Use

Keywords: Data Integration, Application-driven Integration, Data-driven Integration, Web Services, RPC, Publish & Subscribe, Consolidation, Data Warehouse, Data Integration, Service Oriented Architecture , Virtual Data Integration, Query complexity, heterogeneity

The integration problem…

Based on Carlo Batini [13]



What Integration Criteria to Use

Based on Carlo Batini [13]

1. Autonomy, the degree of independence between the different database administrators in their design choices;

2. Relevance of historical data, and consequent need to periodically store new data without deleting the old ones;

3. Query complexity, in terms of amount of data and tables visited and number of operators on them, and consequent time complexity in query execution;

4. Relevance of currency in queries, the need for queries to extract current data;

5. Economic value of integration, the relevance of having integrated information in input for business operational and decisional processes in order to produce effective outputs;

What Integration Criteria to Use

Based on Carlo Batini [13]

6. Volatility of sources, frequency of adding or deleting sources, and frequency of change of source schemas;

7. Relevance of queries w.r.t transactions, relative importance and frequency of queries with respect to changes in data;

8. Management complexity, the effort to be spent in management activities related to databases and hw-sw infrastructures, due to the corresponding complexity of the organizations using the data bases;

9. Costs of heterogeneity, hidden and explicit costs related to business processes that are due to making use of heterogeneous data.

References

[1] Mustafa Jarrar, Anton Deik: The Graph Signature: A Scalable Query Optimization Index for RDF Graph Databases Using Bisimulation and Trace Equivalence Summarization. International Journal on Semantic Web and Information Systems, 11(2), 36-65,. April-June 2015

[2] Mustafa Jarrar, Anton Deik, Bilal Faraj: Ontology-Based Data And Process Governance Framework -The Case Of E-Government Interoperability In Palestine . In pre-proceedings of the IFIP International Symposium on Data-Driven Process Discovery and Analysis (SIMPDA'11). Pages(83-98). 2011.

[3] Mustafa Jarrar and Marios D. Dikaiakos: A Query Formulation Language for the Data Web. The IEEE Transactions on Knowledge and Data Engineering. IEEE Computer Society. Pages(783-798). Volume 24, Number 4, April 2012

[4] Paolo Ceravolo, Chengfei Liu, Mustafa Jarrar, Kai-Uwe Sattler: Special Issue on Querying the Data Web -Novel techniques for querying structured data on the web. The World Wide Web Journal. Volume(14), Issue (5-6). Springer. August 2011. ISSN:1573-1413.

[5] Anton Deik, Bilal Faraj, Ala Hawash, Mustafa Jarrar: Towards Query Optimization for the Data Web - Two Disk-Based algorithms: Trace Equivalence and Bisimilarity. Proceedings of the 3rd Palestinian International Conference on Computer and Information Technology (PICCIT 2010). 2010.

[6] Mustafa Jarrar, Marios D. Dikaiakos: Querying the Data Web: the MashQL Approach. IEEE Internet Computing. Volume 14, No. 3. Pages (58-670). IEEE Computer Society, ISSN 1089-7801. May 2010.

[7] Mustafa Jarrar, Marios D. Dikaiakos: Querying the Data Web: the MashQL Approach. IEEE Internet Computing. Volume 14, No. 3. Pages (58-670). IEEE Computer Society, ISSN 1089-7801. May 2010.Mustafa Jarrar and Marios D. Dikaiakos: A Data Mashup Language for the Data Web . Proceedings of LDOW, WWW'09. ACM. ISSN 1613-0073. (2009).

[8] Mustafa Jarrar and Marios D. Dikaiakos: MashQL: a query-by-diagram topping SPARQL -Towards Semantic Data Mashups. Proceedings of ONISW'08, part of the ACM CiKM conference. ACM. pages (89-96) ISBN 9781605582559.(2008).

[0] Mustafa Jarrar: Towards methodological principles for ontology engineering. PhD Thesis. Vrije Universiteit Brussel. (May 2005)

[10] Mustafa Jarrar, Luk Vervenne, Diana Maynard: HR-Semantics Roadmap- The Semantic challenges and opportunities in the Human Resources domain . Technical Report. The Ontology Outreach Advisory, Belgium. (OOA-HR/2007-08-20/v025). August 2007

[11] Lyndon Nixon, Malgorzata Mochol, Mustafa Jarrar, Stamatia Dasiopoulou, Vasileios Papastathis, and Yiannis Kompatsiaris: Prototypical business use cases. Deliverable D1.1.2 (WP1.1), The Knowledge Web Network of Excellence (NoE) IST-2004-507482, Luxemburg. January 2005.

[12] Peter Spyns, Daniel Oberle, Raphael Volz, Jijuan Zheng, Mustafa Jarrar, York Sure, Rudi Studer, and Robert Meersman: OntoWeb- a Semantic Web Community Portal. Proceedings of the 4th International Conference on Practical Aspects of Knowledge Management (PAKM 2002). Pages (189-200). LNCS 2569, Springer. ISBN: 3540003142. December 2002.

[13] Carlo Batini: Course on Data Integration. BZU IT Summer School 2011.

[14] Stefano Spaccapietra: Information Integration. Presentation at the IFIP Academy. Porto Alegre. 2005.

[15] Chris Bizer: The Emerging Web of Linked Data. Presentation at SRI International, Artificial Intelligence Center. Menlo Park, USA. 2009.