

Semi-automatic Generation of Active Ontologies from Web Forms

Martin Blersch, Mathias Landhäußer, and Thomas Mayer

Institute for Program Structures and Data Organization (IPD) – Programming Systems, Chair Tichy



"Create an appointment with Peter on Friday."



SUN	MON	TUE	WED	THU	FRI	SAT
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	19	
20	21	22	23	24	25	26
27	28	29	30	31		

www.free-printable-calendar.com

How can we

automatically
add new features to

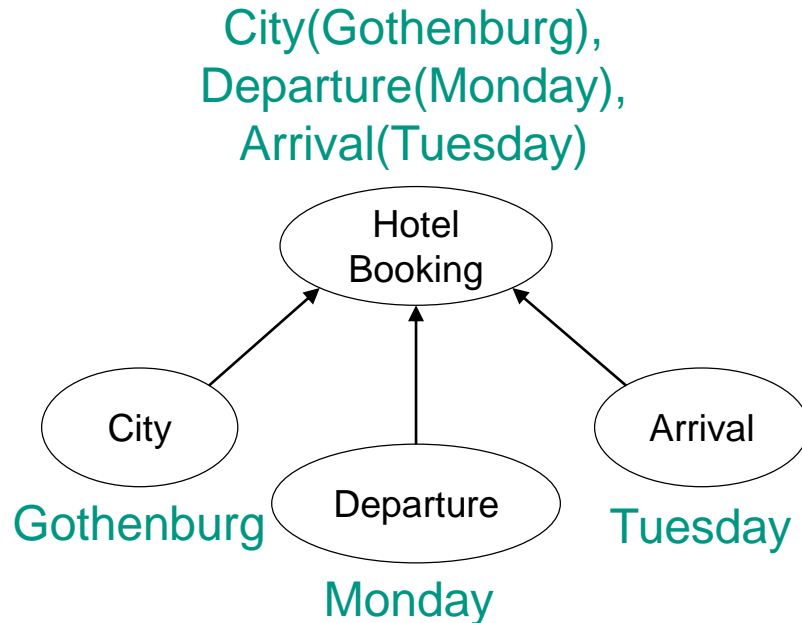
intelligent assistants ?

EASIER: A Framework to Connect Intelligent Assistants with Arbitrary Web Forms



- Generates 65% of the software automatically
- Correctly answers 70% of the queries with the generated software

Active Ontologies / Active Semantic Network



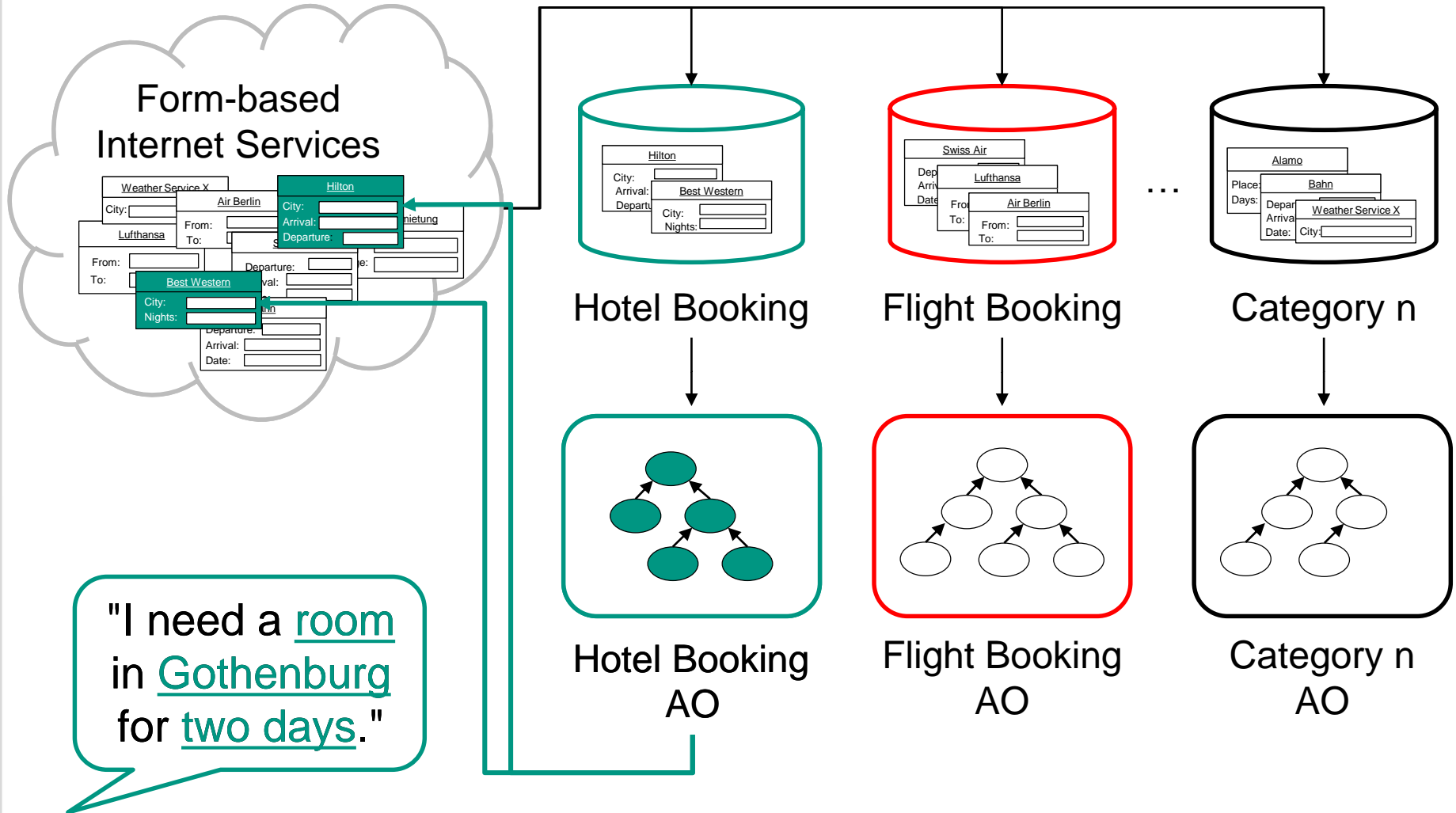
"I need a hotel room
In Gothenburg from
Monday to Tuesday."

- Combine the modelling of domain knowledge with an execution environment
- Different node types
 - Leaf nodes
 - Non-terminal nodes
- Bottom-up processing of utterances

- Supporting new features requires extending the ontologies or even building new ones

→ Manual & labor-intensive steps

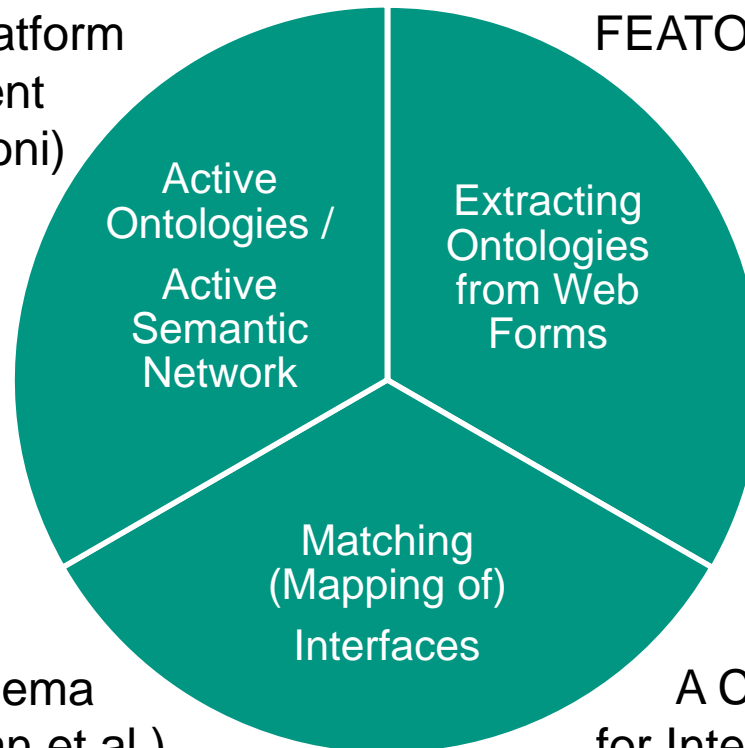
EASIER – Big Picture



Related Work

Active: A Unified Platform
for Building Intelligent
Applications (Guzzoni)

FEATON: Builds Ontologies from
Forms semi-automatically
(Berlanga et al.)



OntoBuilder:
Extracts Ontologies
from Web Forms
(Gal & Roitman et al.)

Cupid: Generic Schema
Matching (Madhavan et al.)

A Clustering based Approach
for Interface Matching (Wu et al.)

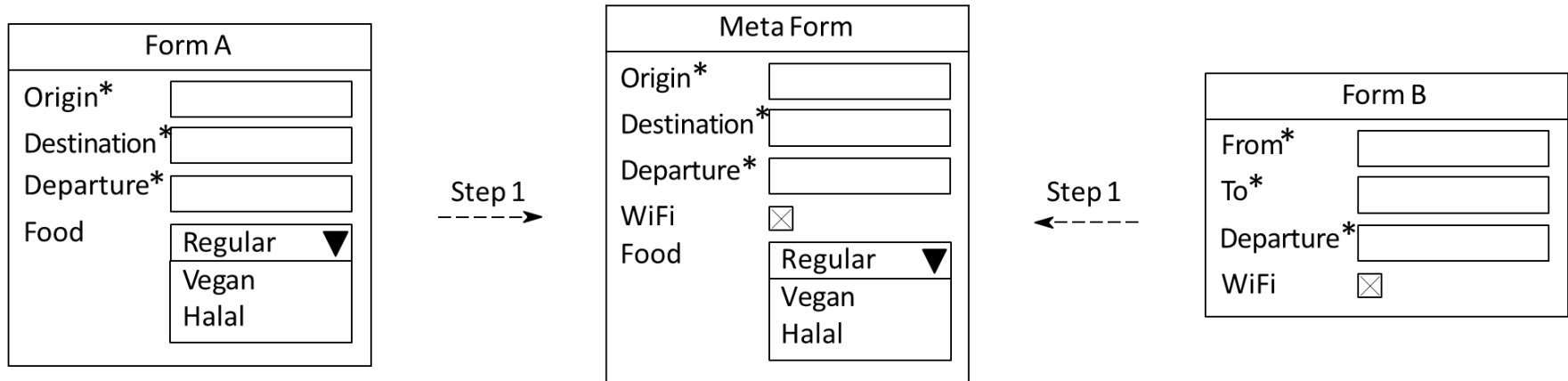
WISE: Automatic Integrator for
Web Search Interfaces (He et al.)

Automatically Creating Active Ontologies: Overview

Form A	
Origin*	<input type="text"/>
Destination*	<input type="text"/>
Departure*	<input type="text"/>
Food	<div>Regular ▼ Vegan Halal</div>

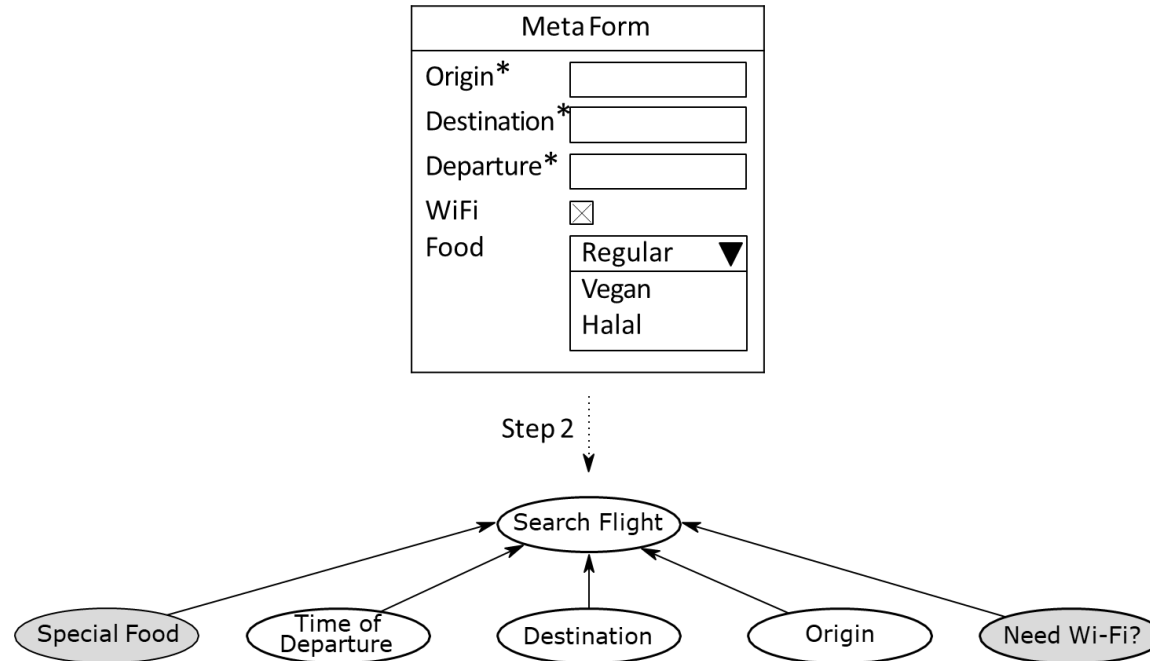
Form B	
From*	<input type="text"/>
To*	<input type="text"/>
Departure*	<input type="text"/>
WiFi	<input type="checkbox"/>

Automatically Creating Active Ontologies: Grouping Related Form Elements



- Grouping similar form elements
 - Uses linguistic and structural similarities
 - Hierarchical Clustering of similar form elements
- Generation of meta form elements (global objects)
 - Merge form elements

Automatically Creating Active Ontologies: Deriving the Ontology



- One Active Ontology for each meta form (i.e., one per category)
 - Value range unclear → ask developer
 - Select AO node type

Evaluation

- Three evaluation questions
 - How good is the field matching?
 - What is the degree of automation?
 - Are the generated AOs capable to answer user queries?

- Data Source
 - 58 web forms from the UIUC Web Integration Repository
 - Three categories: airfare, automobile, and book search

- Queries
 - 61 queries for the airfare domain
 - 40 subjects (20 of them are native English speaker)

Evaluation: Field Matching

Category	Precision	Recall	$F_{0,5}$ Measure
Airfare	90.6	21.0	54.4
Automobile	90.6	37.3	70.5
Book	98.4	46.4	80.4

- Highest precision for the book search domain
- Recall of clustering must be improved

Evaluation: Degree of Automation

Category	Nodes			
	Total	Manual	Autom.	Autom. [%]
Airfare	126	29	97	77%
Automobile	41	23	18	44%
Book	49	24	25	51%
Total	216	76	140	65%

- Automatically generated 77% of the elements needed for the airfare domain
- Needed the help of developers in only 35% of the cases

Evaluation: Query Answering

- How many query elements were correctly identified by the sensor nodes?
 - 61 queries from airfare domain
 - E.g. "Book a flight from Frankfurt to Paris."
 - Post-processing of given values
 - Expected: origin, destination, departure
 - Ask user for missing information
 - Results
 - Recall: 75%
 - Queries Completely recognized
 - Only mandatory information: 77.4%
 - Mandatory and optional information: 35.5%

Evaluation: Query Answering (2)

Field	Correct	Wrong	Missing
Origin	191	9	0
Destination	185	15	0
Departure	124	24	52
Total	500	48	52

■ How many queries were correctly identified by the sensor nodes?

- 26 complete queries (out of 61 queries)
- 10 analyzed
- 20 web forms (airfare domain)

■ Results

- 7 queries were recognized correctly
- 2 were not recognized (EASIER asked the user)
- 1 was recognized incorrectly

Conclusion and Future Work

- EASIER automates the process of building AOs
 - Automatically generates 65% of the AO's sensor nodes
 - High precision in field matching (90.6 - 98.4%)

- Queries
 - Correctly answers 70% of the queries
 - Asks for missing information

- Future Work
 - Improve field matching performance (HTML5, ARIA, ...)
 - Integrate complex field mappings
 - Better domain knowledge (Wikipedia, Cyc)

References

- Berlanga, Rafael, Ernesto Jimenez-Ruiz, Victoria Nebot, und Ismael Sanz. „Faeton: Form analysis and extraction tool for ontology construction“. *International Journal of Computer Applications in Technology* 39, Nr. 4 (2010): 224–33.
- Guzzoni, Didier. „Active: A Unified Platform for Building Intelligent Applications“. PhD Thesis, École Polytechnique Fédérale De Lausanne, 2008.
http://biblion.epfl.ch/EPFL/theses/2008/3990/3990_abs.pdf.
- He, Hai, Weiyi Meng, Clement Yu, und Zonghuan Wu. „Automatic Integration of Web Search Interfaces with WISE-Integrator“. *The VLDB Journal* 13, Nr. 3 (1. September 2004): 256–73.
<https://doi.org/10.1007/s00778-004-0126-4>.

References

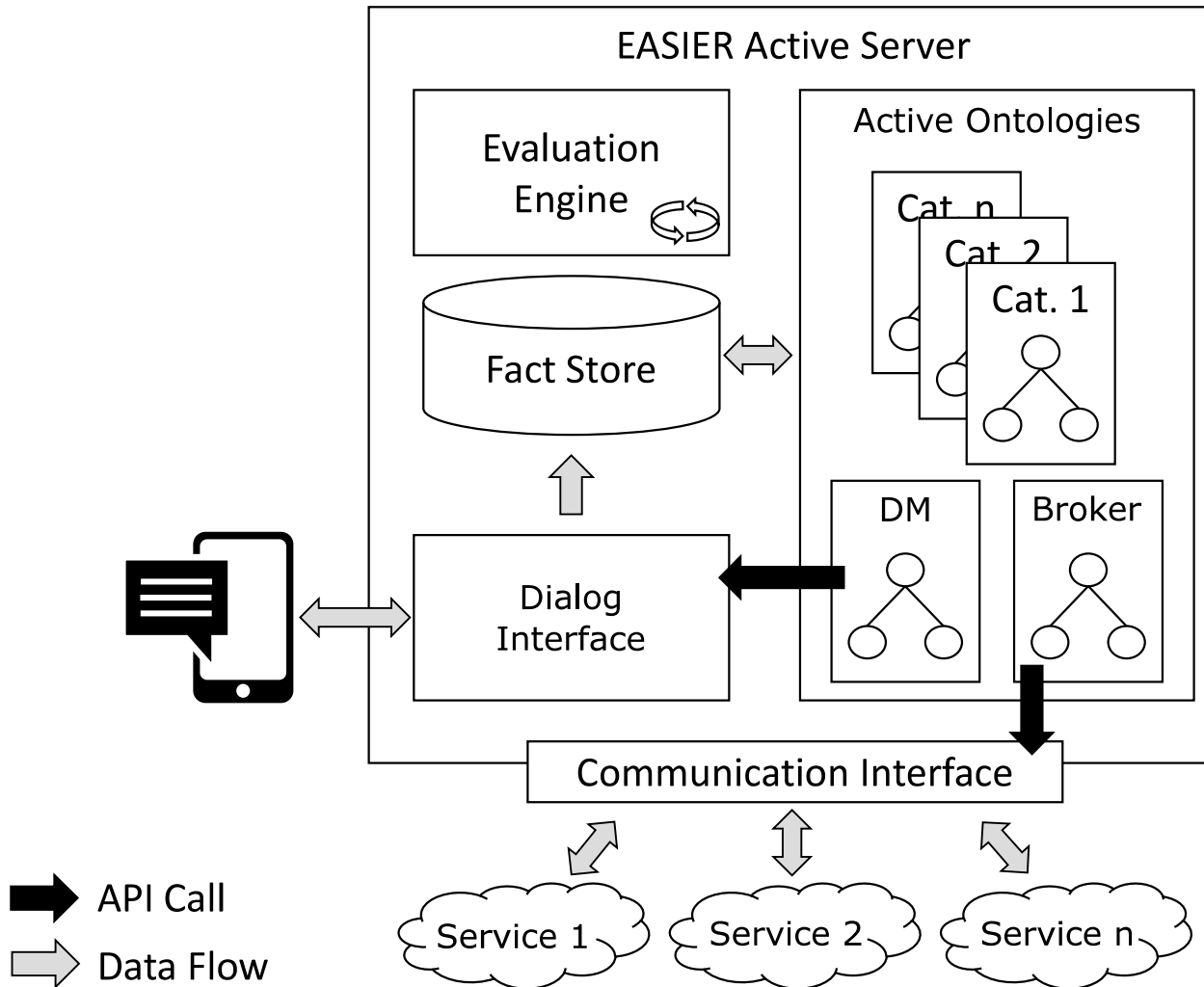
- He, Hai, Weiyi Meng, Clement Yu, und Zonghuan Wu. „Wise-Integrator: An automatic integrator of web search interfaces for e-commerce“. In *Proceedings of the 29th international conference on Very large data bases-Volume 29*, 357–368. VLDB Endowment, 2003.
<http://dl.acm.org/citation.cfm?id=1315483>.
- Madhavan, Jayant, Philip A. Bernstein, und Erhard Rahm. „Generic schema matching with cupid“. Technical Report. Microsoft Research, August 2001. <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-2001-58.pdf>.
- Roitman, Haggai, und Avigdor Gal. „Ontobuilder: Fully automatic extraction and consolidation of ontologies from web sources using sequence semantics“. In *Current Trends in Database Technology--EDBT 2006*, 573–76. Springer, 2006.

References

- Wu, Wensheng, Clement Yu, AnHai Doan, und Weiyi Meng. „An Interactive Clustering-based Approach to Integrating Source Query Interfaces on the Deep Web“. In *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, 95–106. SIGMOD '04. Paris, France: ACM, 2004.
<https://doi.org/10.1145/1007568.1007582>.

BACKUP

The EASIER Active Server Architecture



Inspired by [Guz08]