

Foundations and Algorithms to Compute the Provenance of Missing Data

Melanie Herschel
(proposed thesis topic)

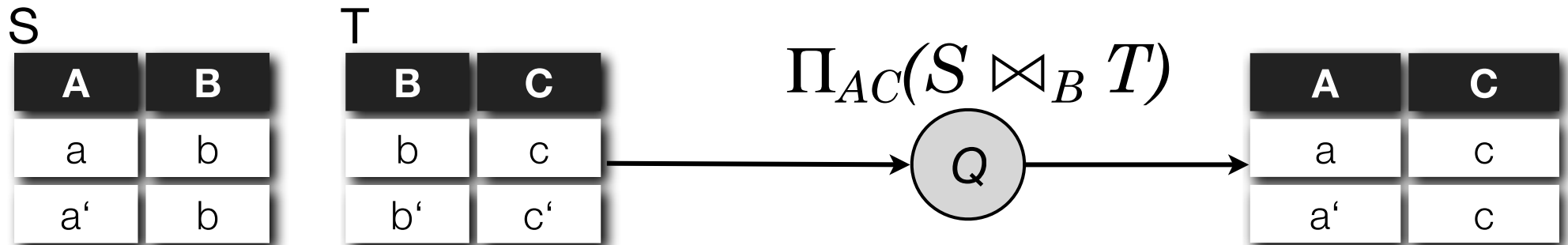
OAK Seminar
June 22, 2012
Dampierre, France

Explaining Missing Answers

Why is some data not in the result of a query Q ?

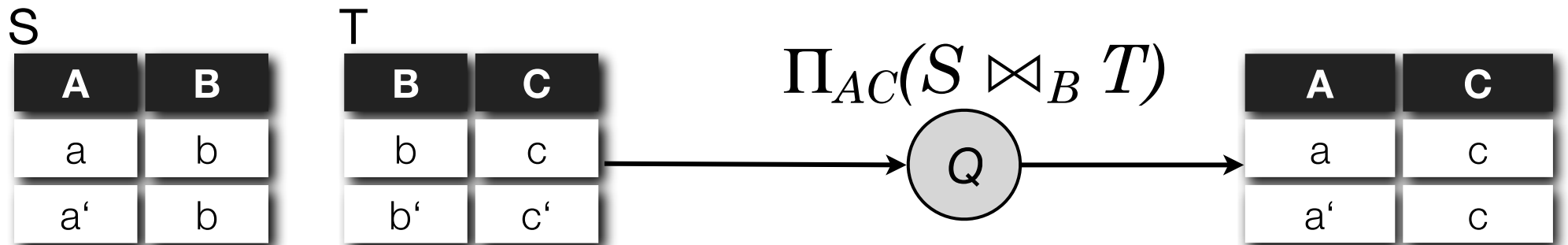
Explaining Missing Answers

Why is some data not in the result of a query Q ?



Explaining Missing Answers

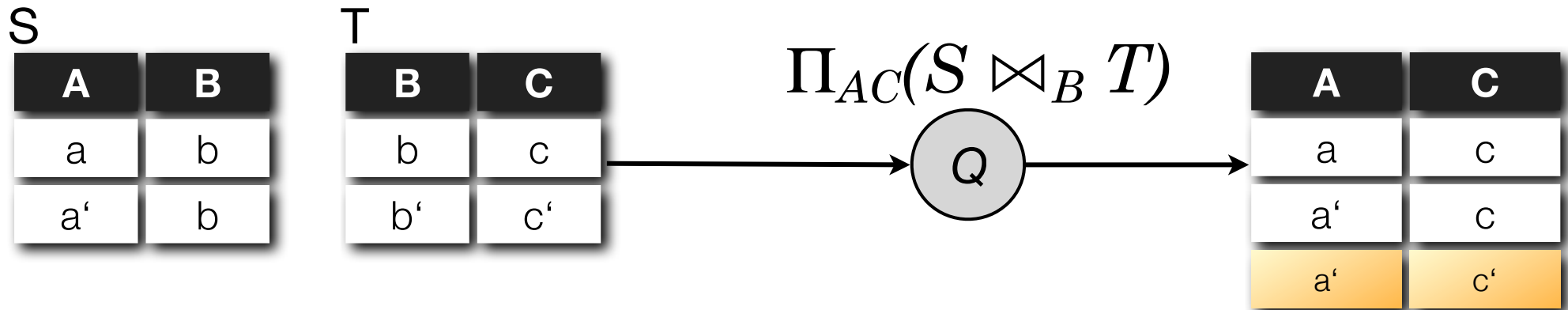
Why is some data not in the result of a query Q?



Why is (a', c') not in the output?

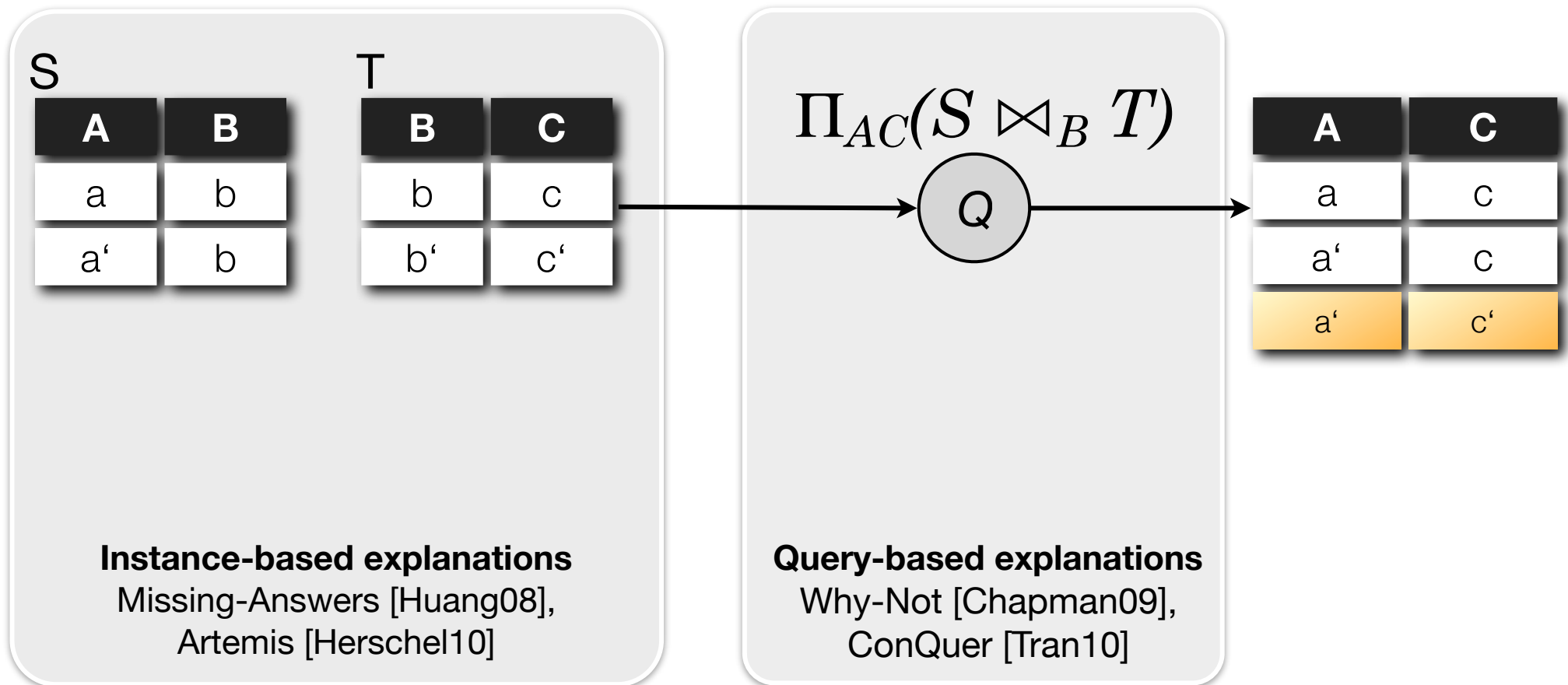
Explaining Missing Answers

Why is some data not in the result of a query Q ?



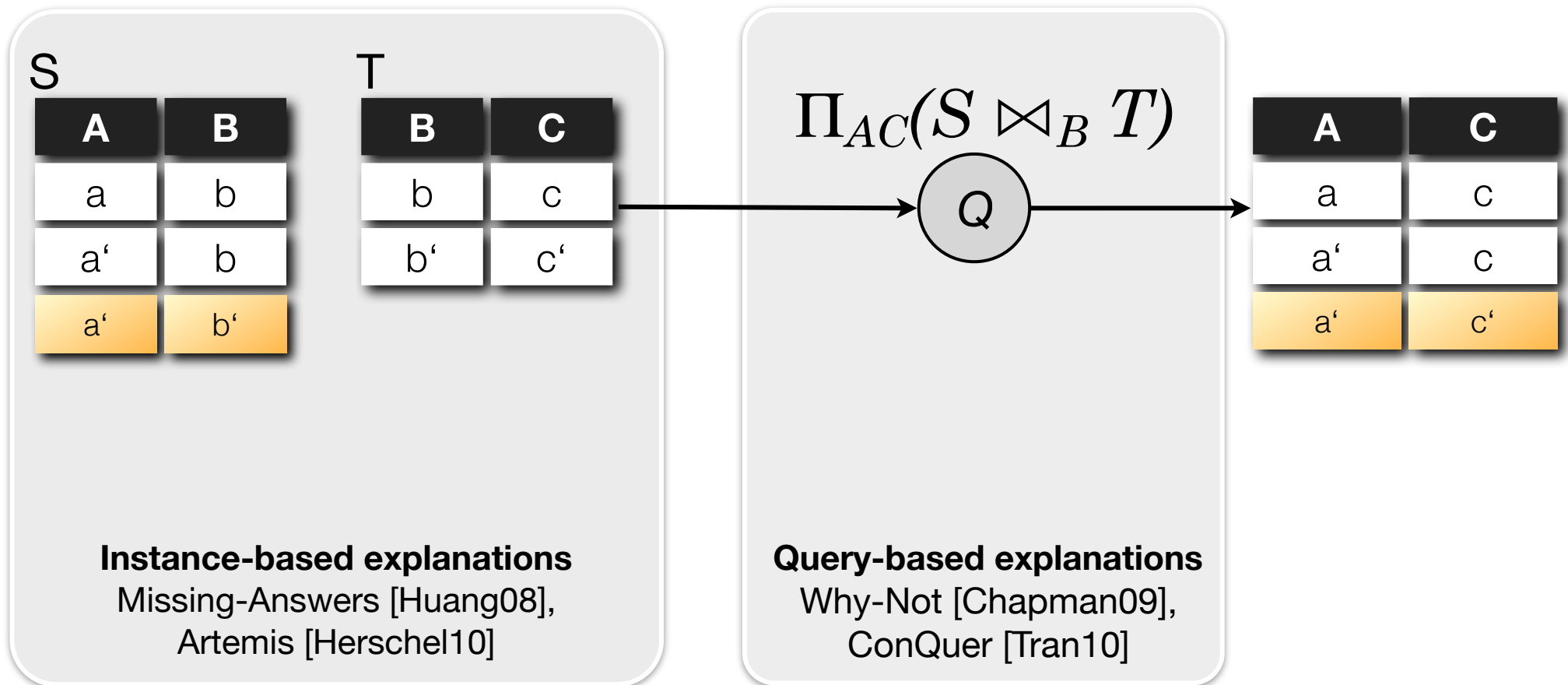
Explaining Missing Answers

Why is some data not in the result of a query Q?



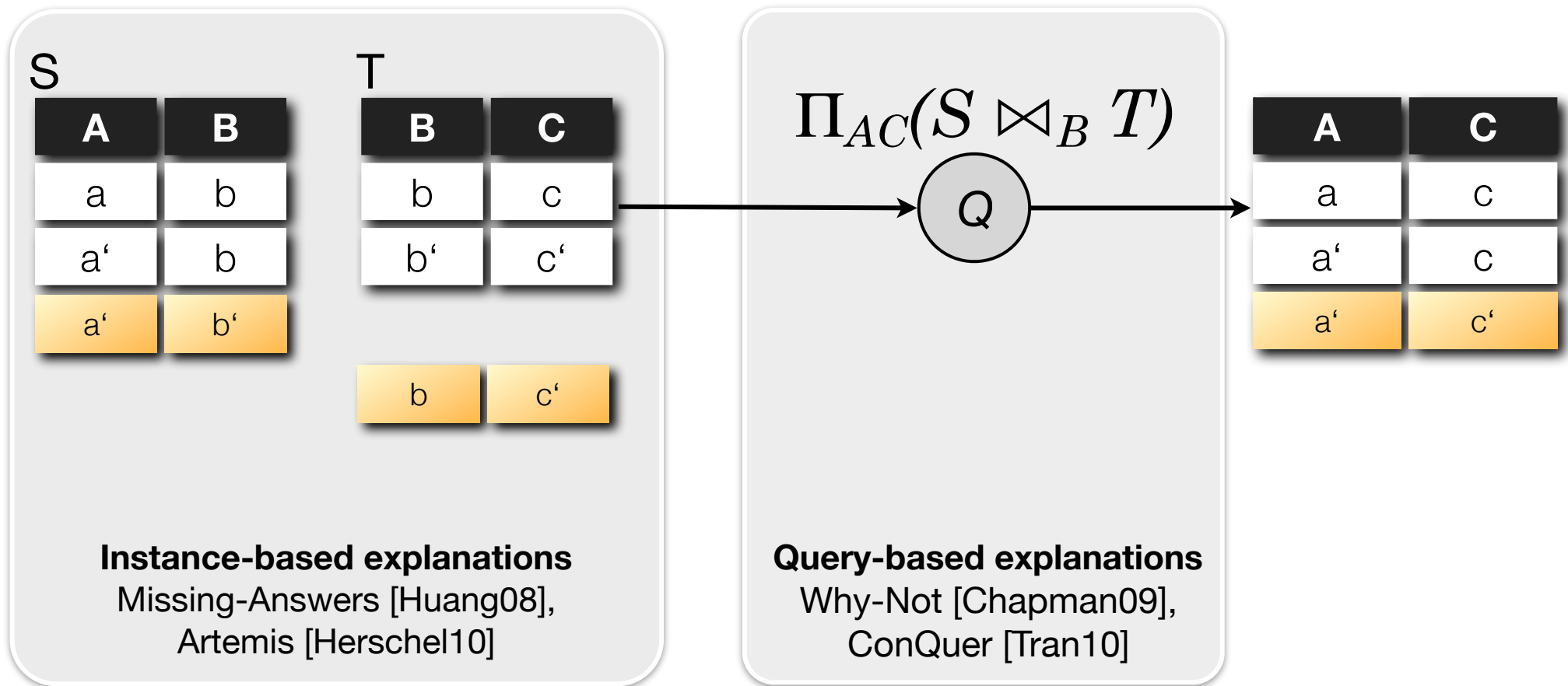
Explaining Missing Answers

Why is some data not in the result of a query Q?



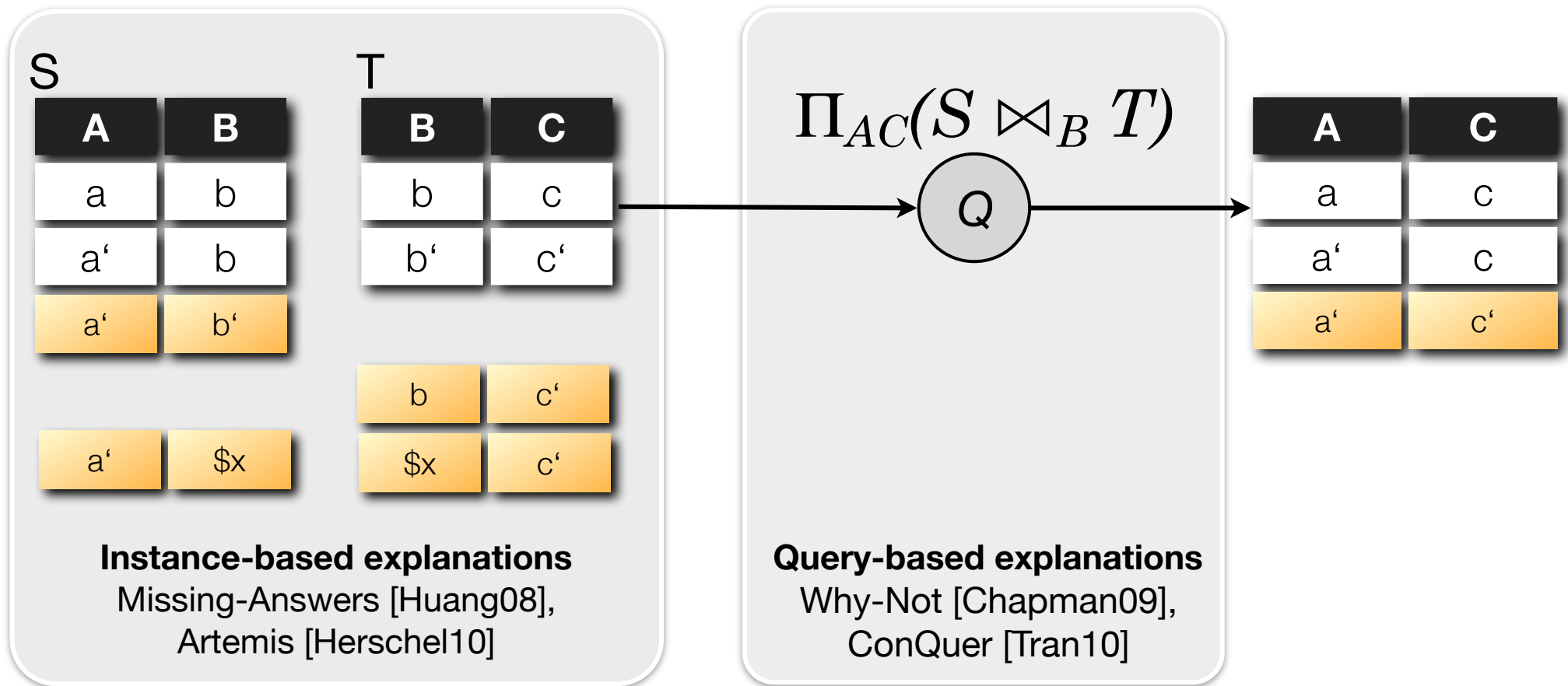
Explaining Missing Answers

Why is some data not in the result of a query Q ?



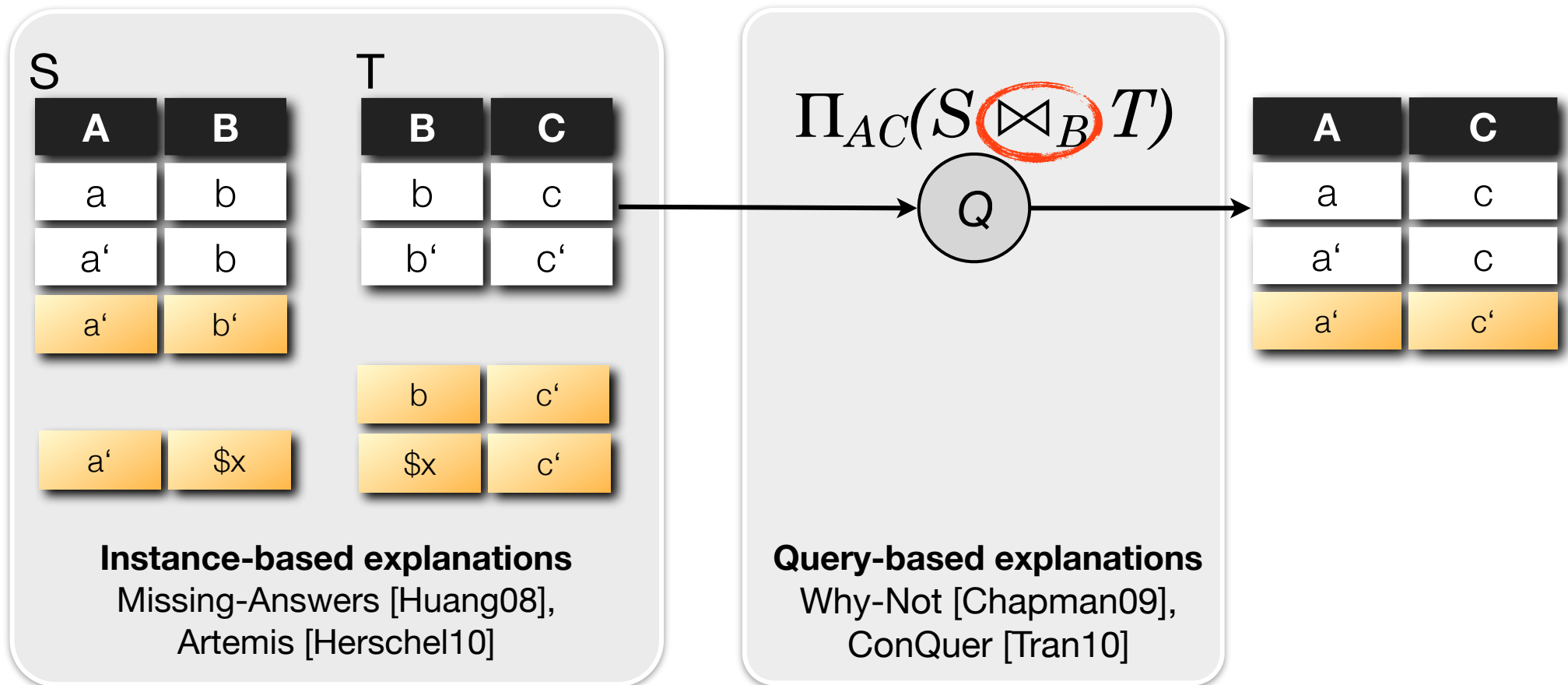
Explaining Missing Answers

Why is some data not in the result of a query Q?



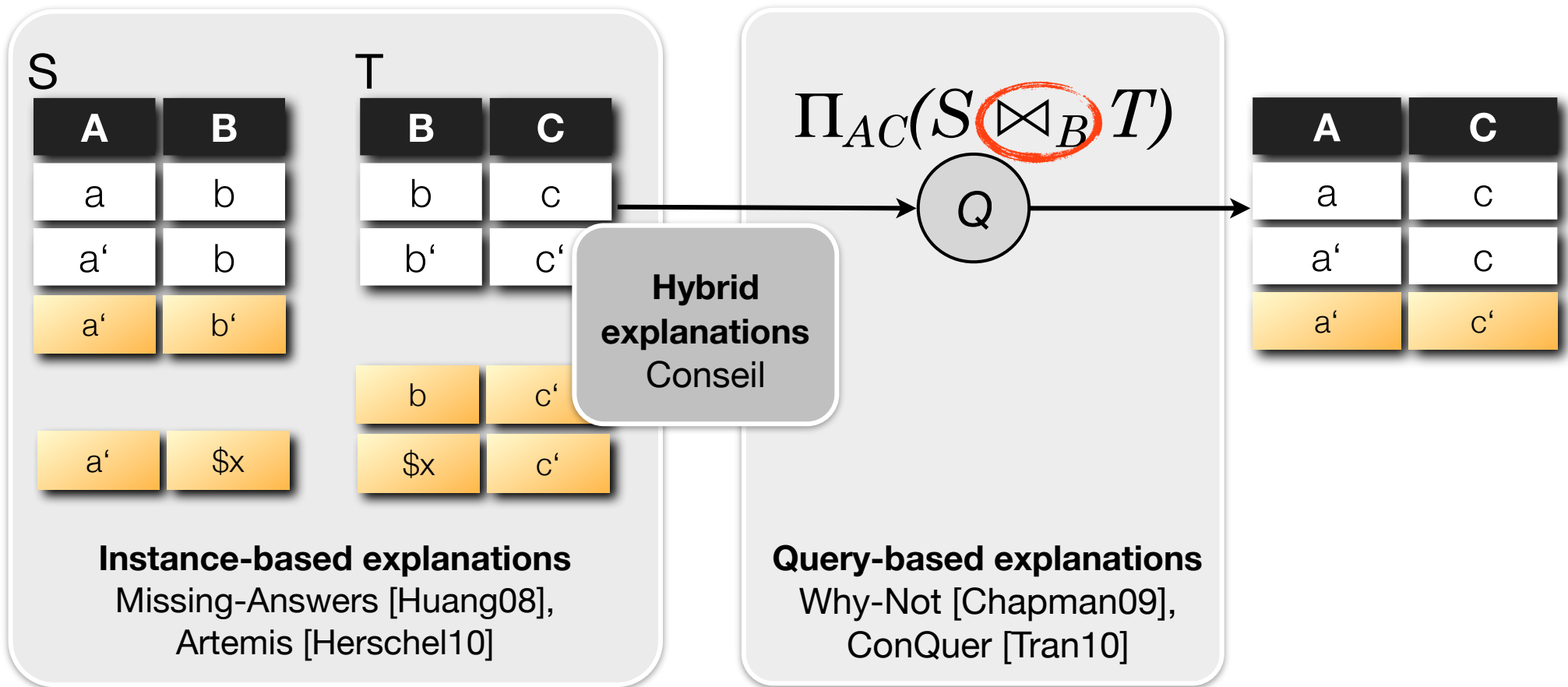
Explaining Missing Answers

Why is some data not in the result of a query Q?



Explaining Missing Answers

Why is some data not in the result of a query Q?

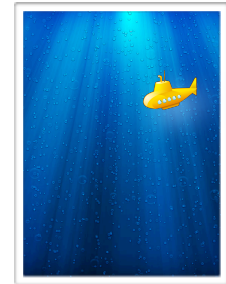


Shortcomings so far

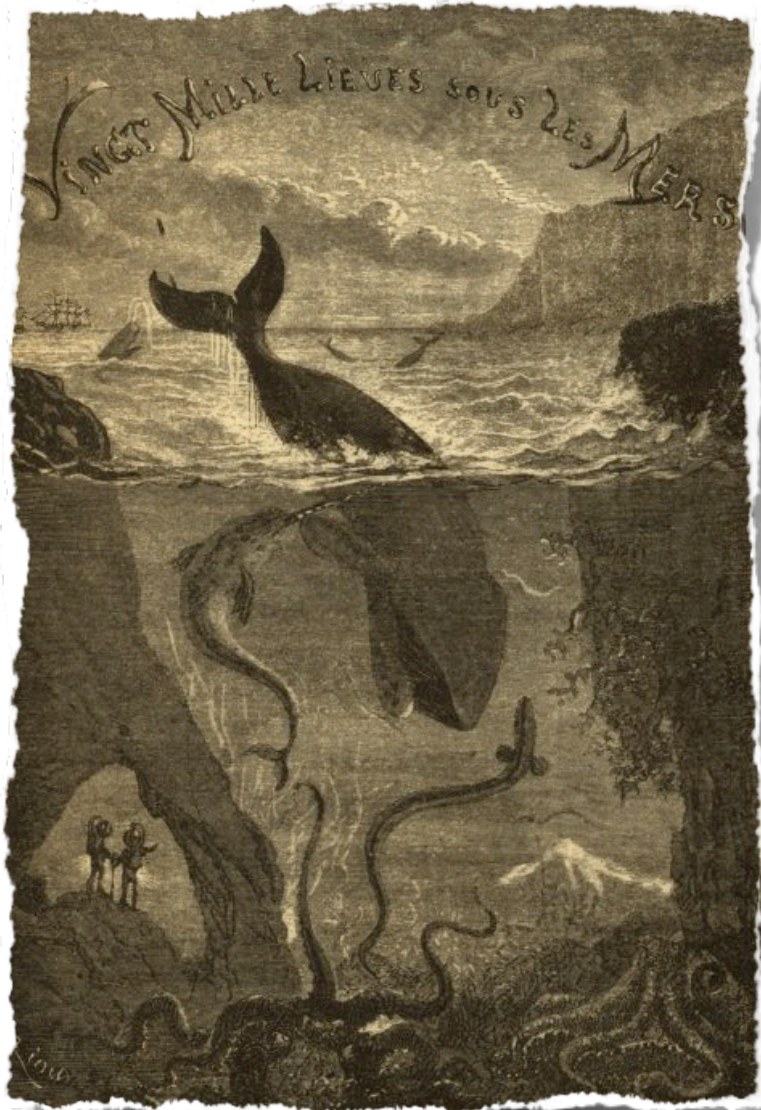
- No grounded **theory** so far
 - Semantics vary
 - Ad-hoc choice of supported (SQL) queries
 - Relationships between explanations only unclear
- Limitation to **subsets of SQL**
- No efficient / scalable **algorithms**

Objectives

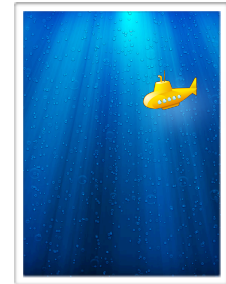
- Development of a **theoretical framework**
 - Formalization of missing-data provenance
 - Identification of interesting properties
 - Problem analysis for different query classes
- Definition of **efficient and effective algorithms**
 - For different types of explanations (instance-based, query-based, hybrid)
 - Use for instance summaries and approximations
- Experimental **validation**
 - In terms of efficiency
 - In terms of usability w.r.t. to our goal of using explanations to analyze and debug complex data transformations in the context of Nautilus.



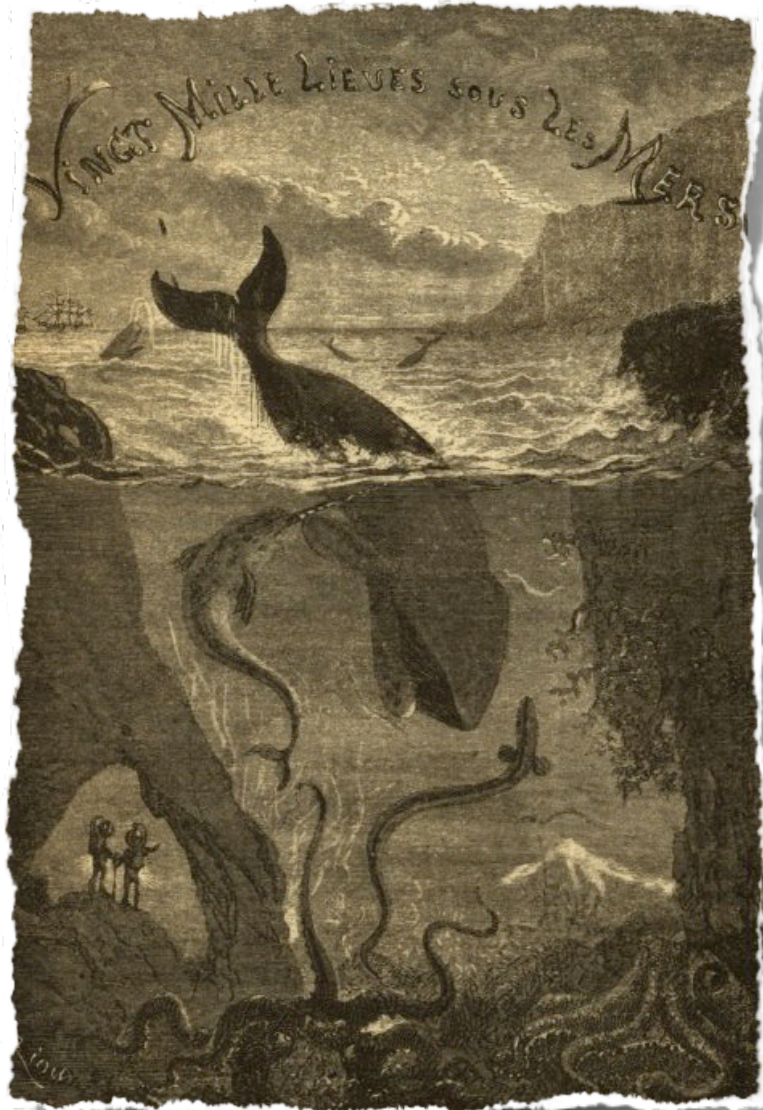
What is Nautilus?



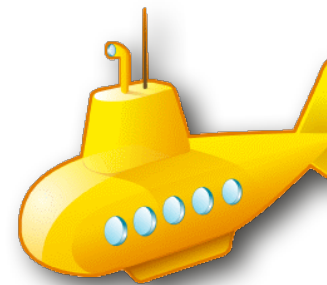
“The **deepest parts of the ocean** are totally unknown to us[...] **What goes on** in those distant depths? What creatures **inhabit**, or could inhabit, those regions twelve or fifteen miles beneath the surface of the water? It's almost beyond **conjecture**” Jules Verne, 20,000 Leagues under the Sea, Chapter 2.

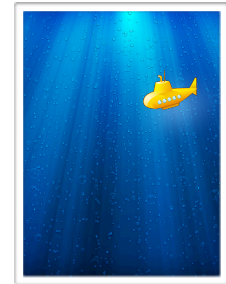


What is Nautilus?



“The **deepest parts of the ocean** are totally unknown to us[...] **What goes on** in those distant depths? What creatures **inhabit**, or could inhabit, those regions twelve or fifteen miles beneath the surface of the water? It's almost beyond **conjecture**” Jules Verne, 20,000 Leagues under the Sea, Chapter 2.



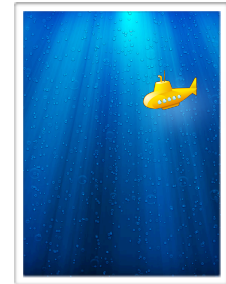


What is Nautilus?



“The **deepest parts of the ocean** are totally unknown to us[...] **What goes on** in those distant depths? What creatures **inhabit**, or could inhabit, those regions twelve or fifteen miles beneath the surface of the water? It's almost beyond **conjecture**” Jules Verne, 20,000 Leagues under the Sea, Chapter 2.





What is Nautilus?



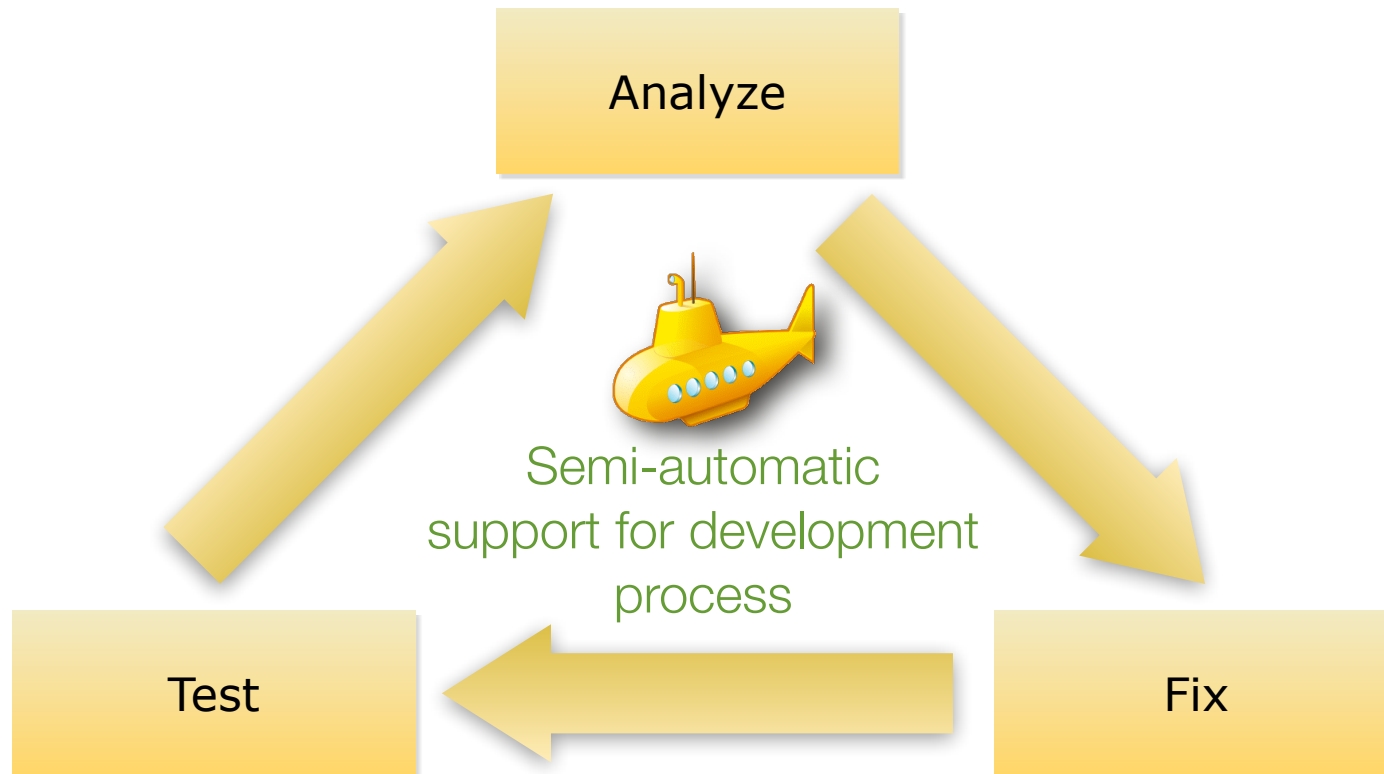
What happens within transformation?

What data?

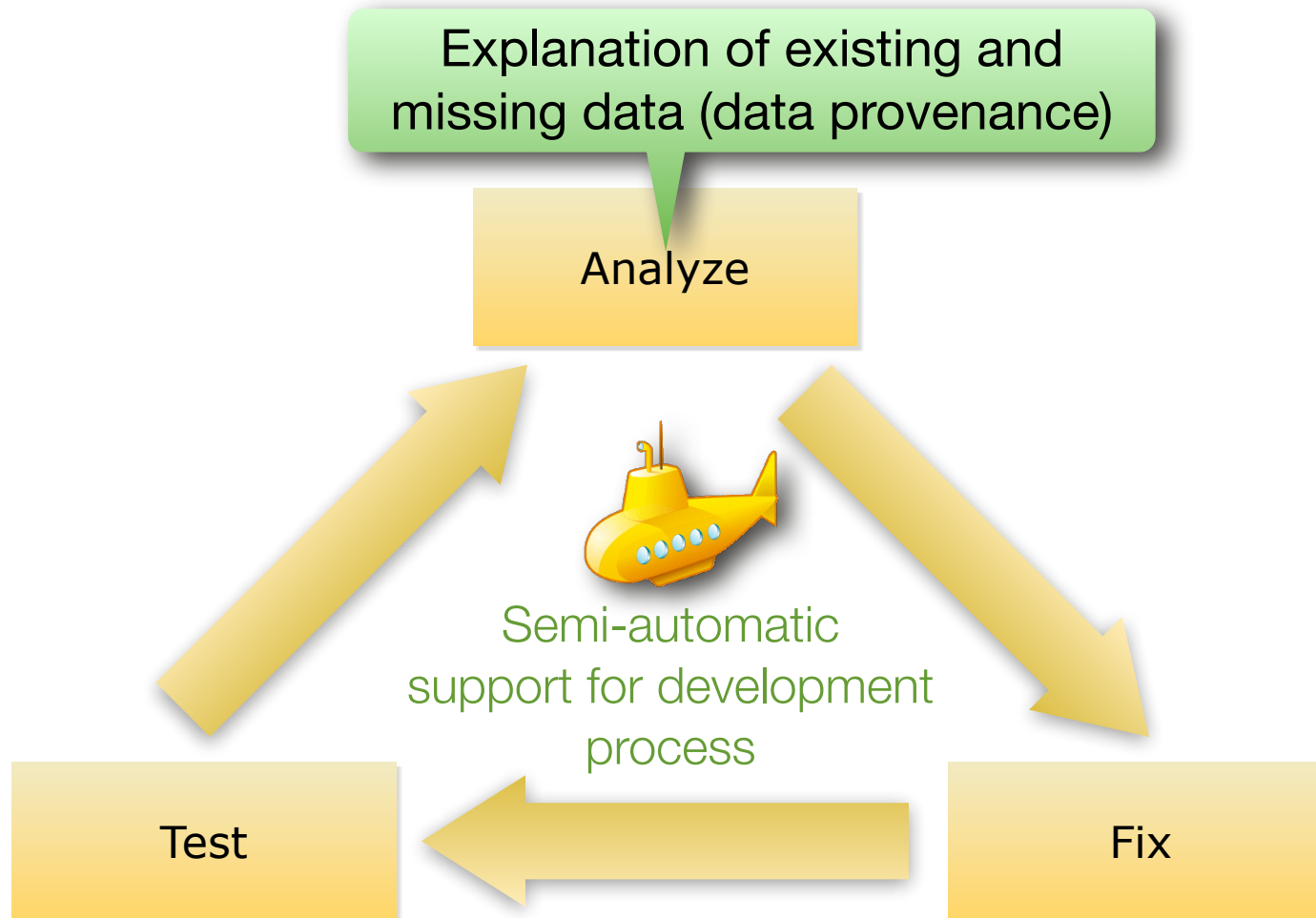
How is data combined?



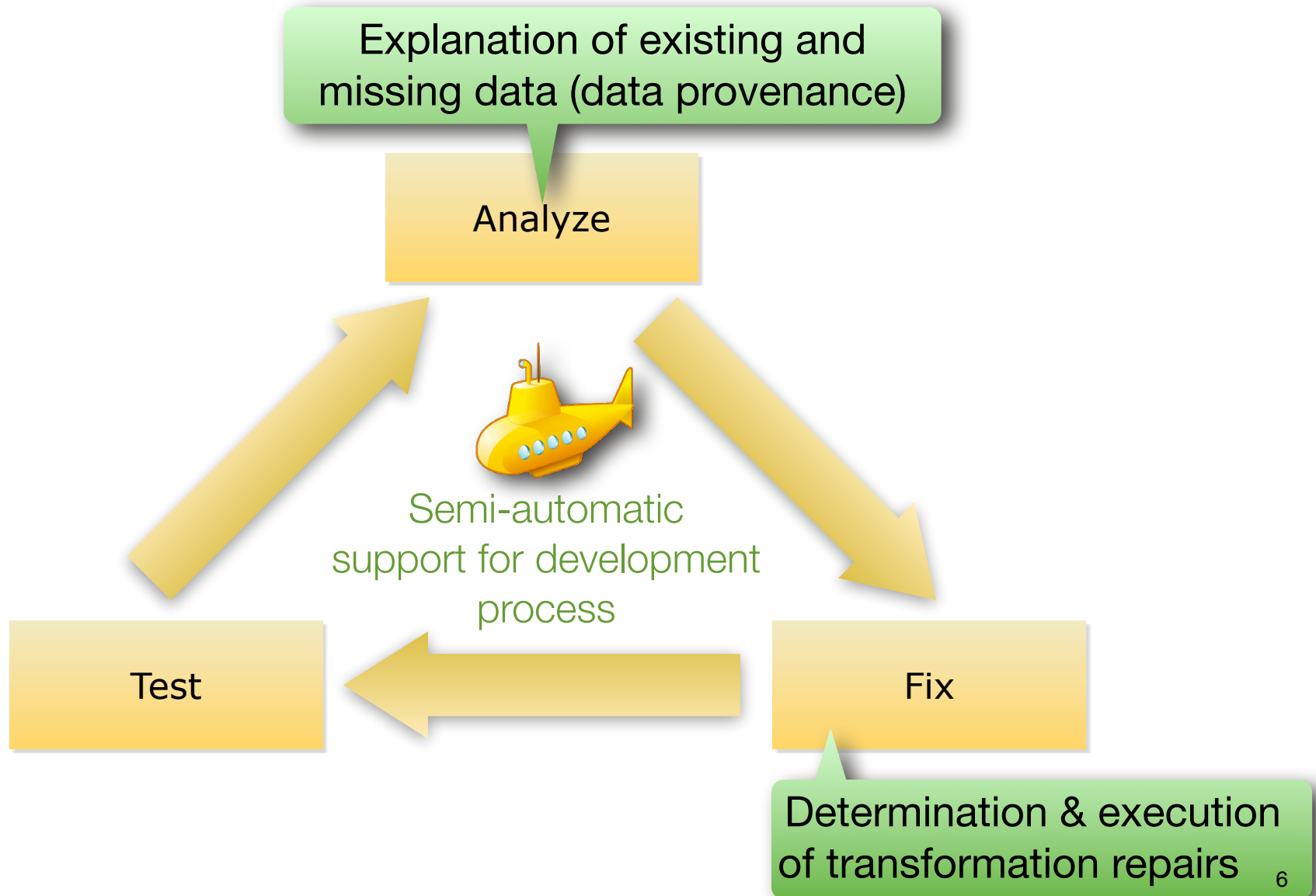
Manual vs. Semi-Automatic Process



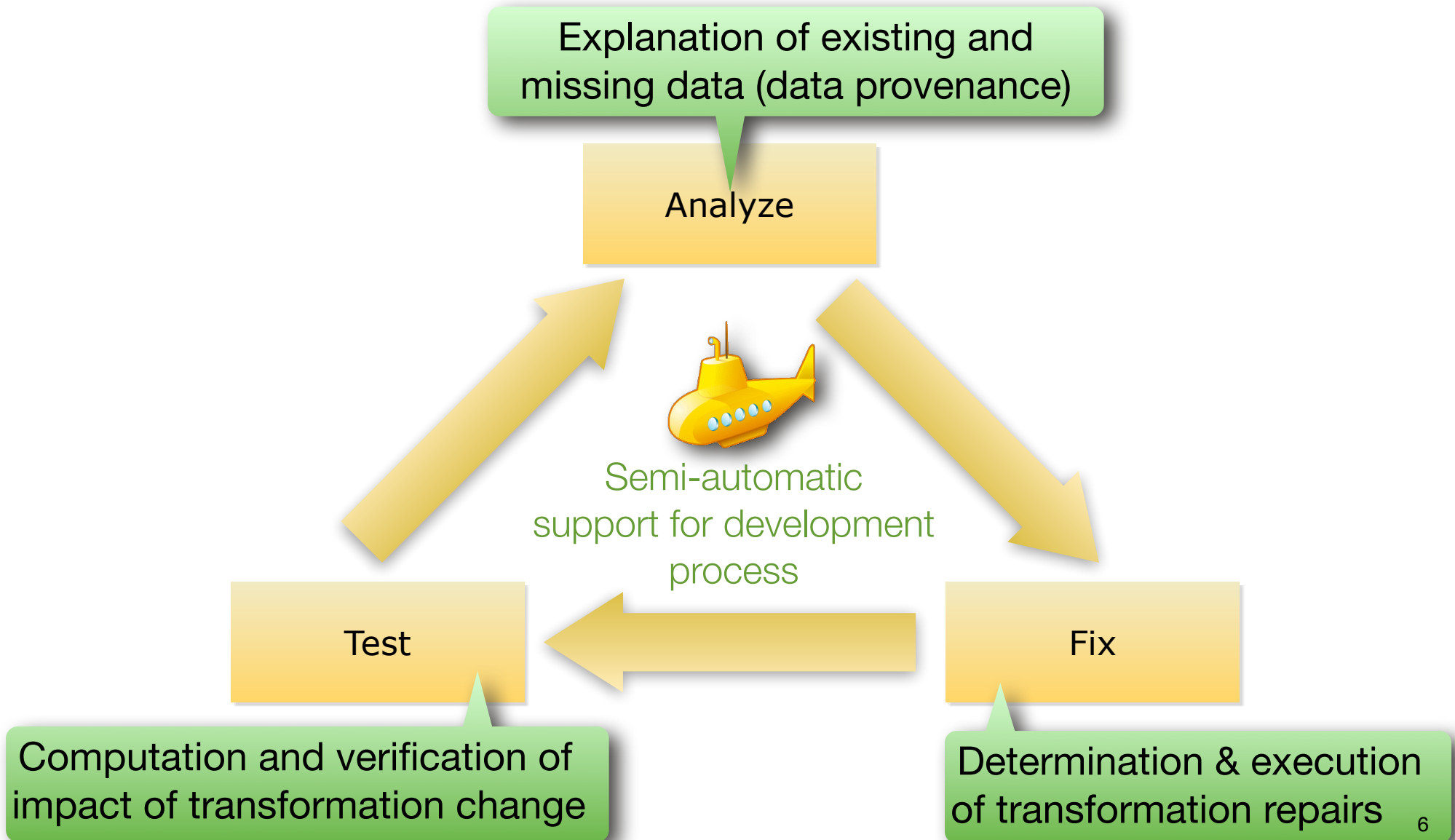
Manual vs. Semi-Automatic Process



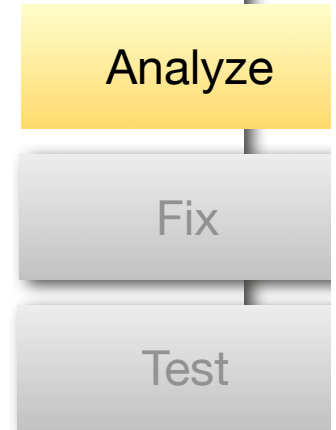
Manual vs. Semi-Automatic Process



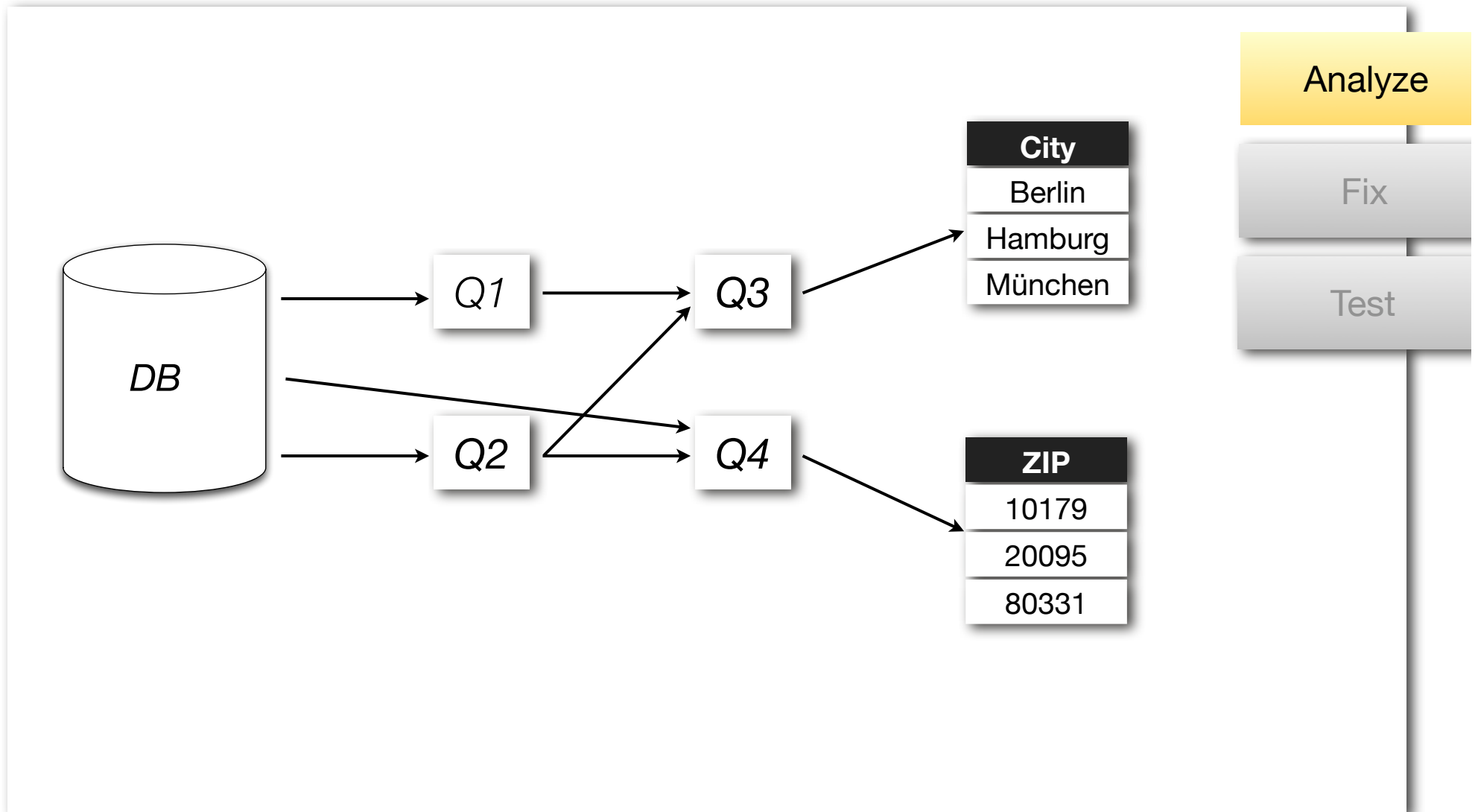
Manual vs. Semi-Automatic Process



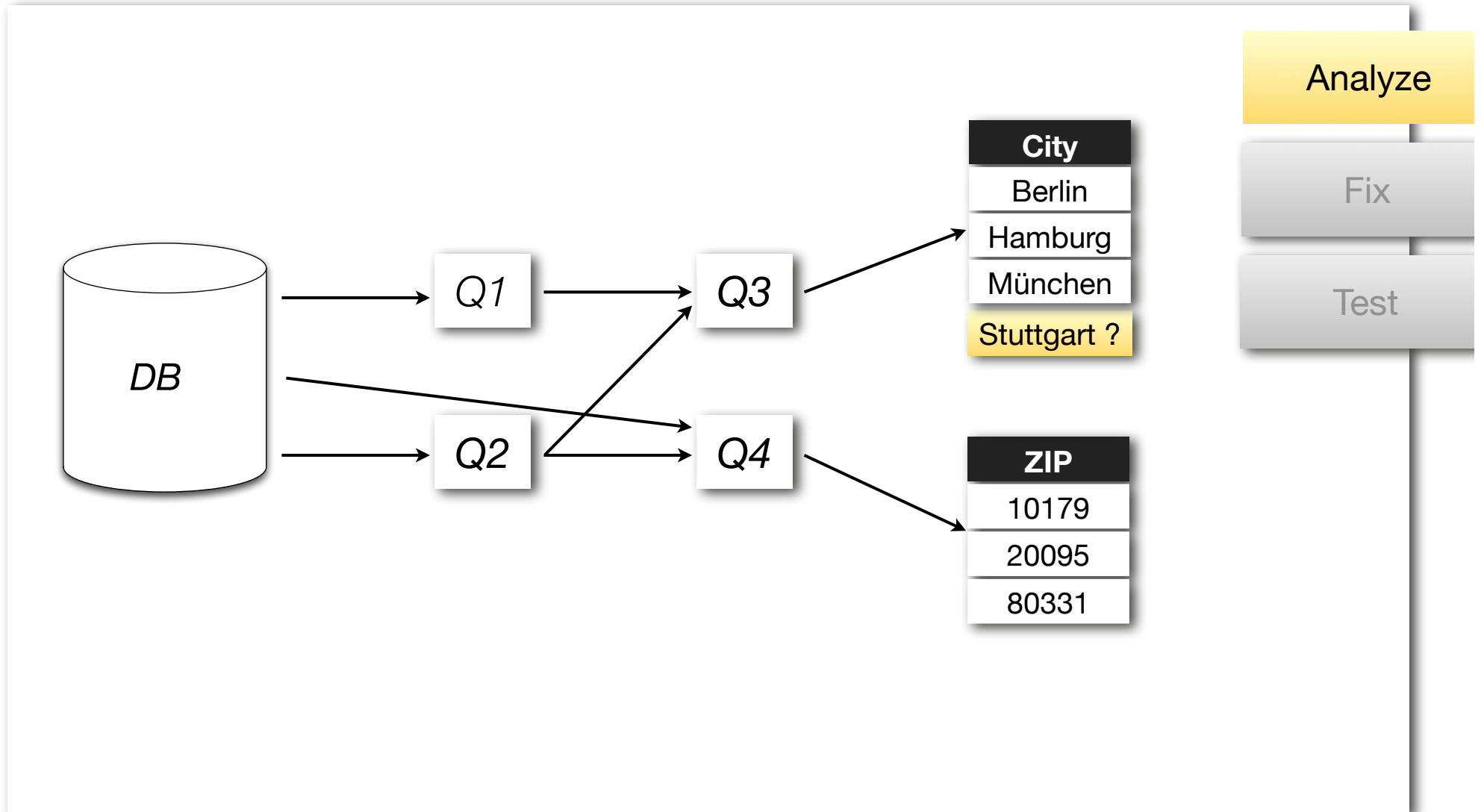
Sample Workflow



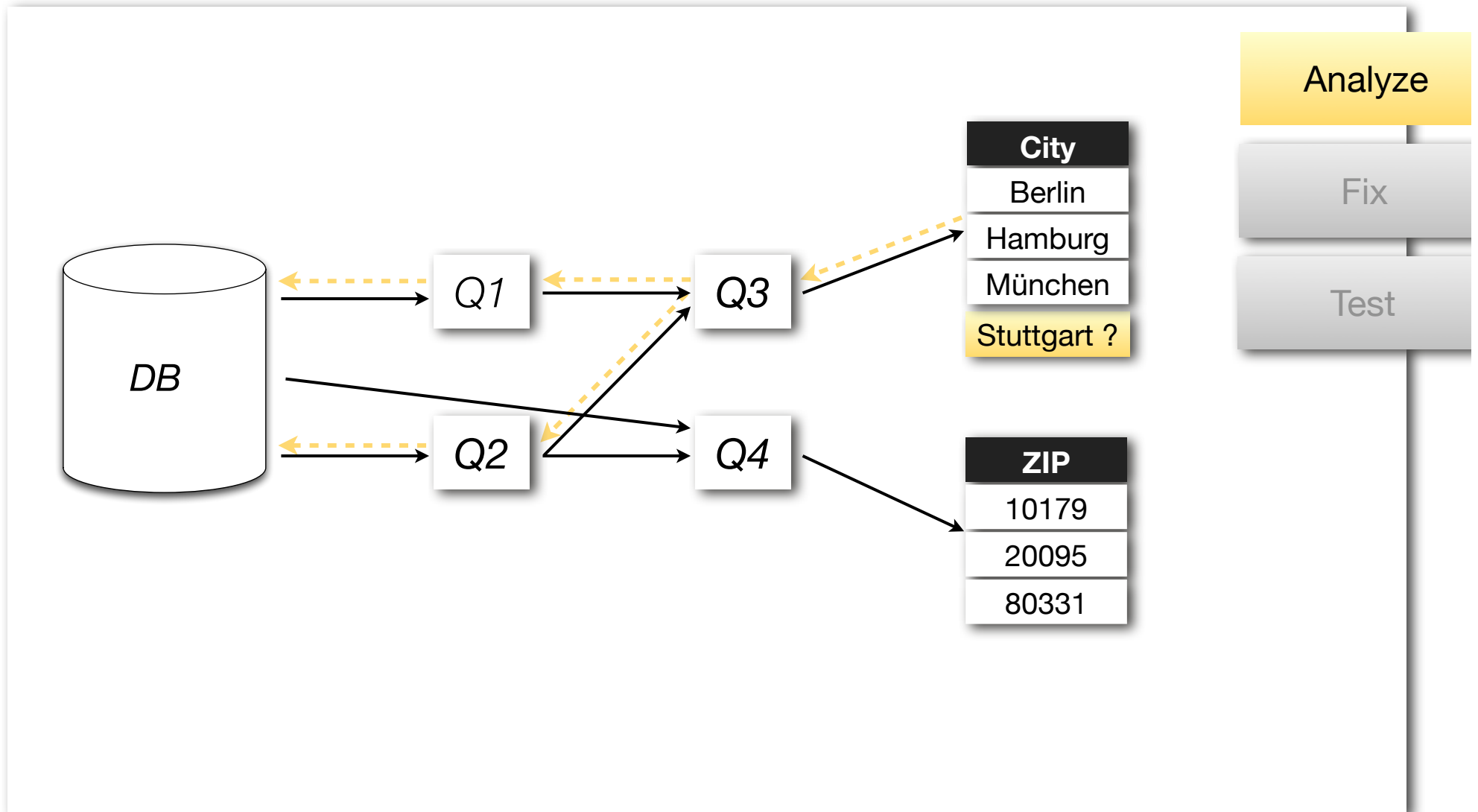
Sample Workflow



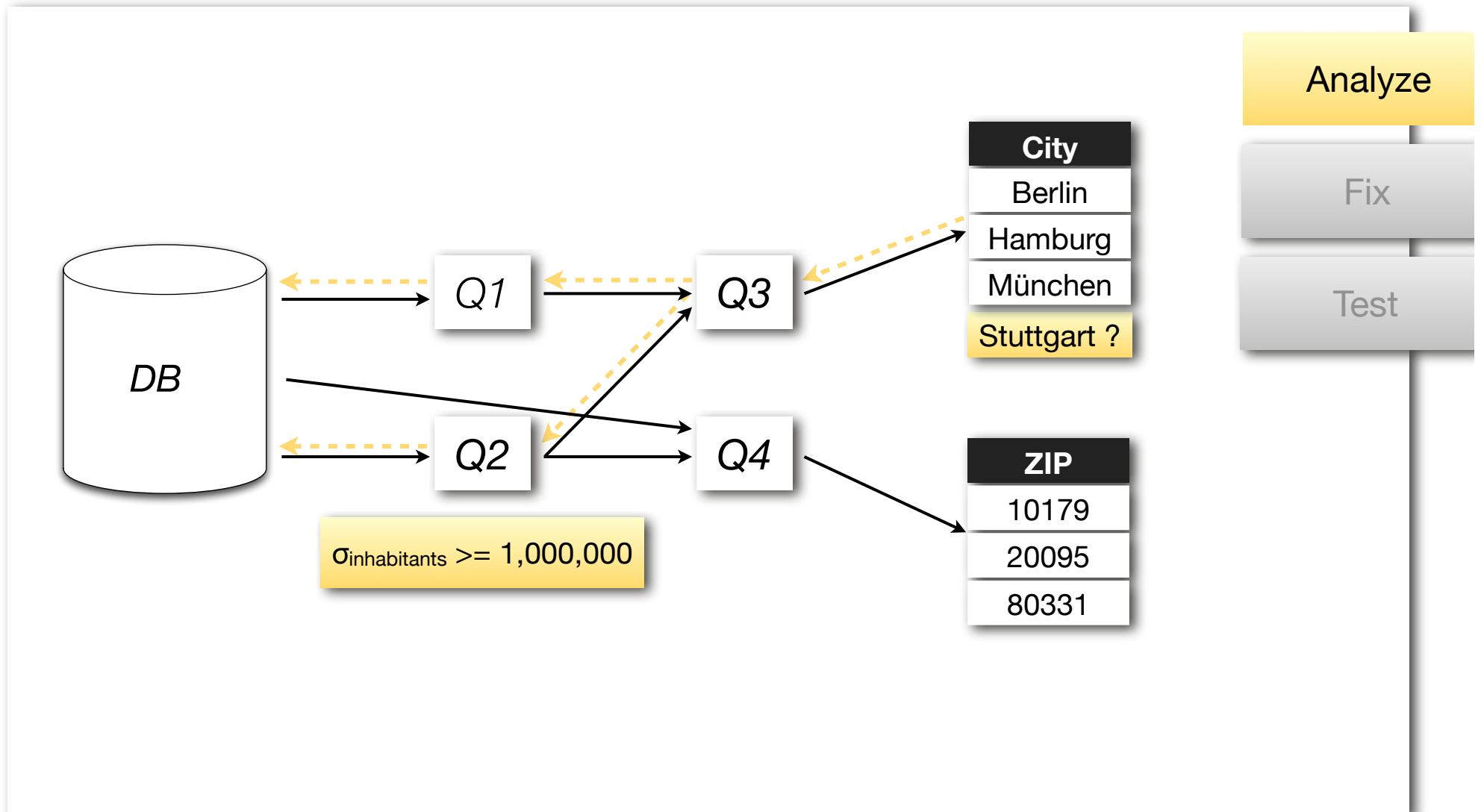
Sample Workflow



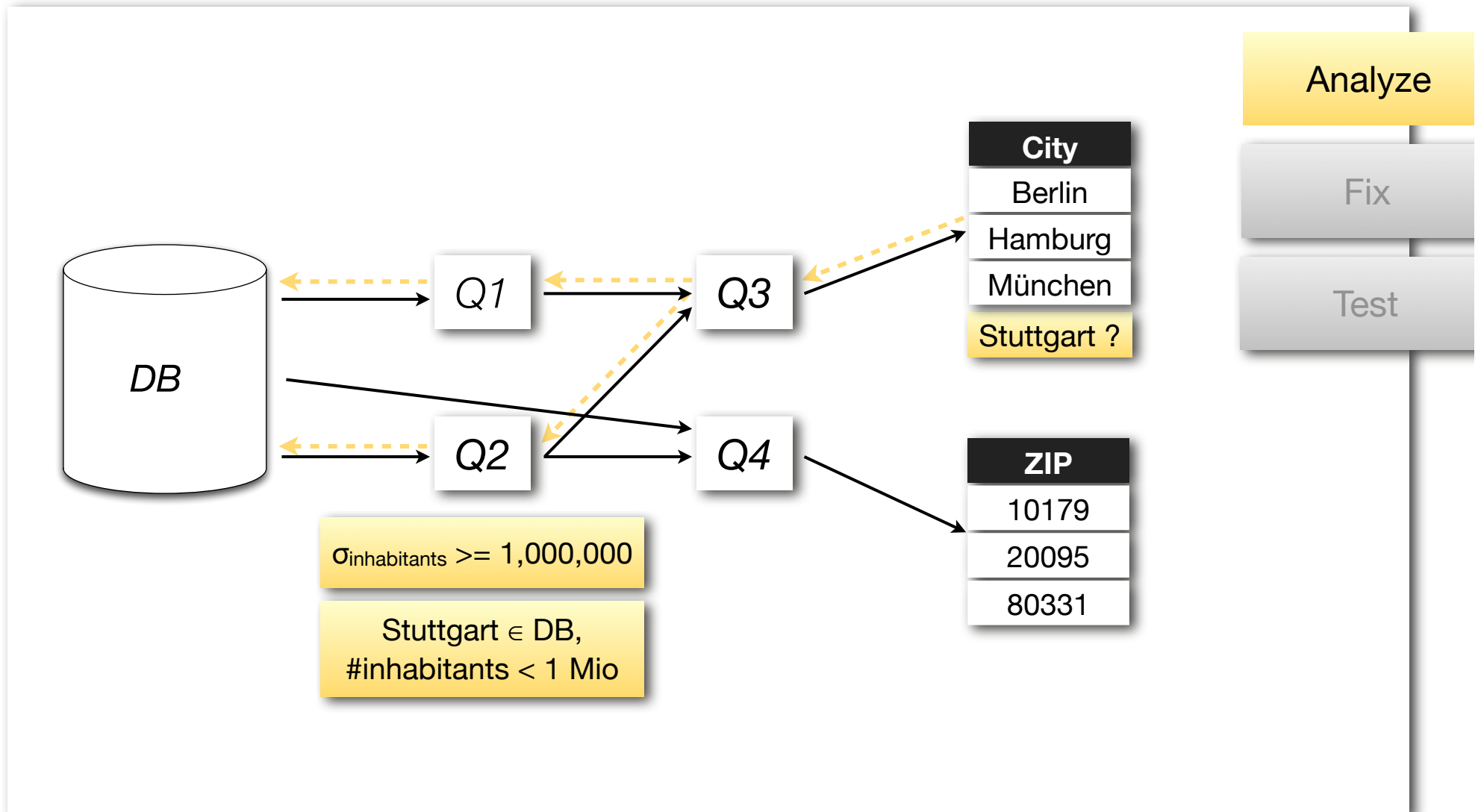
Sample Workflow



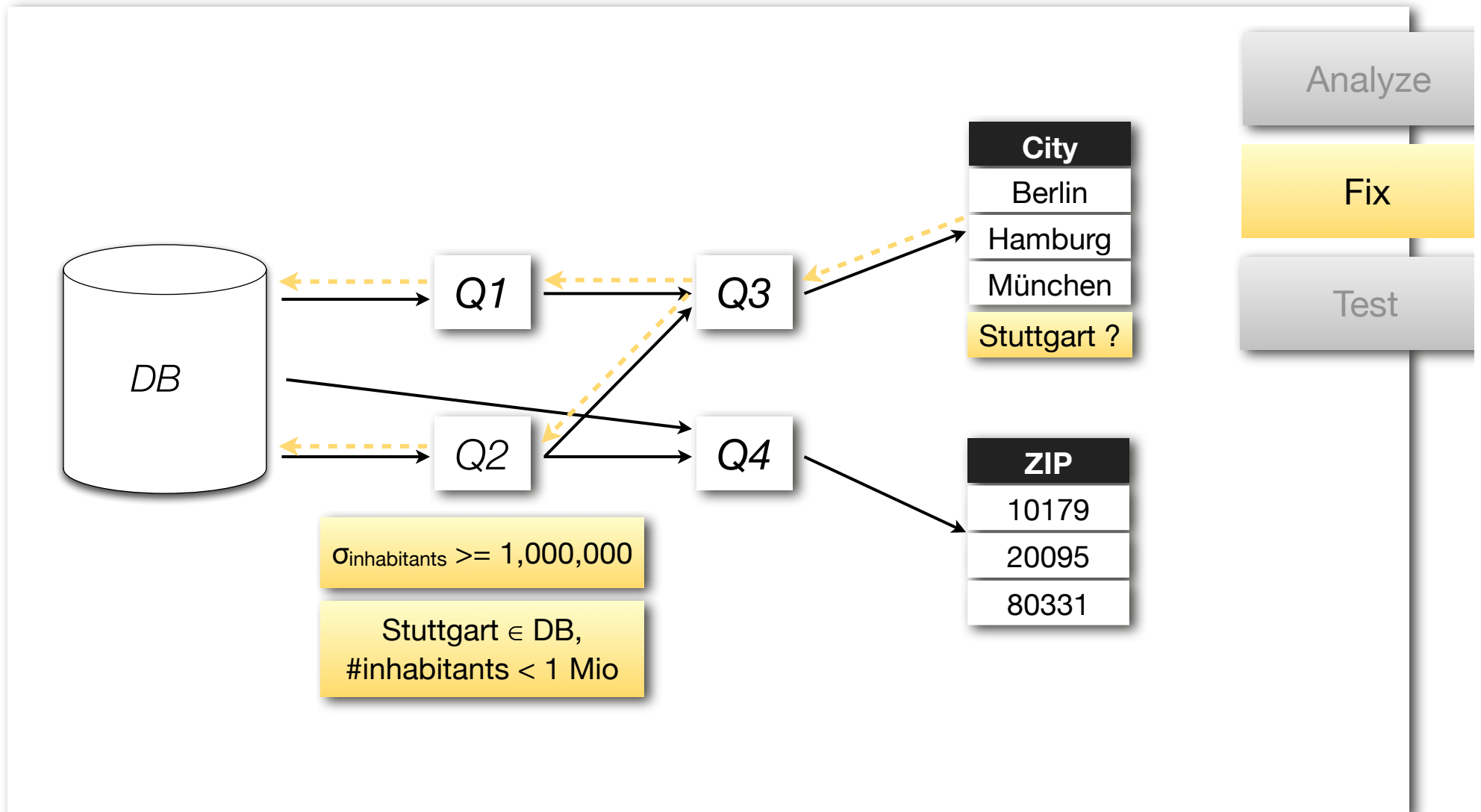
Sample Workflow



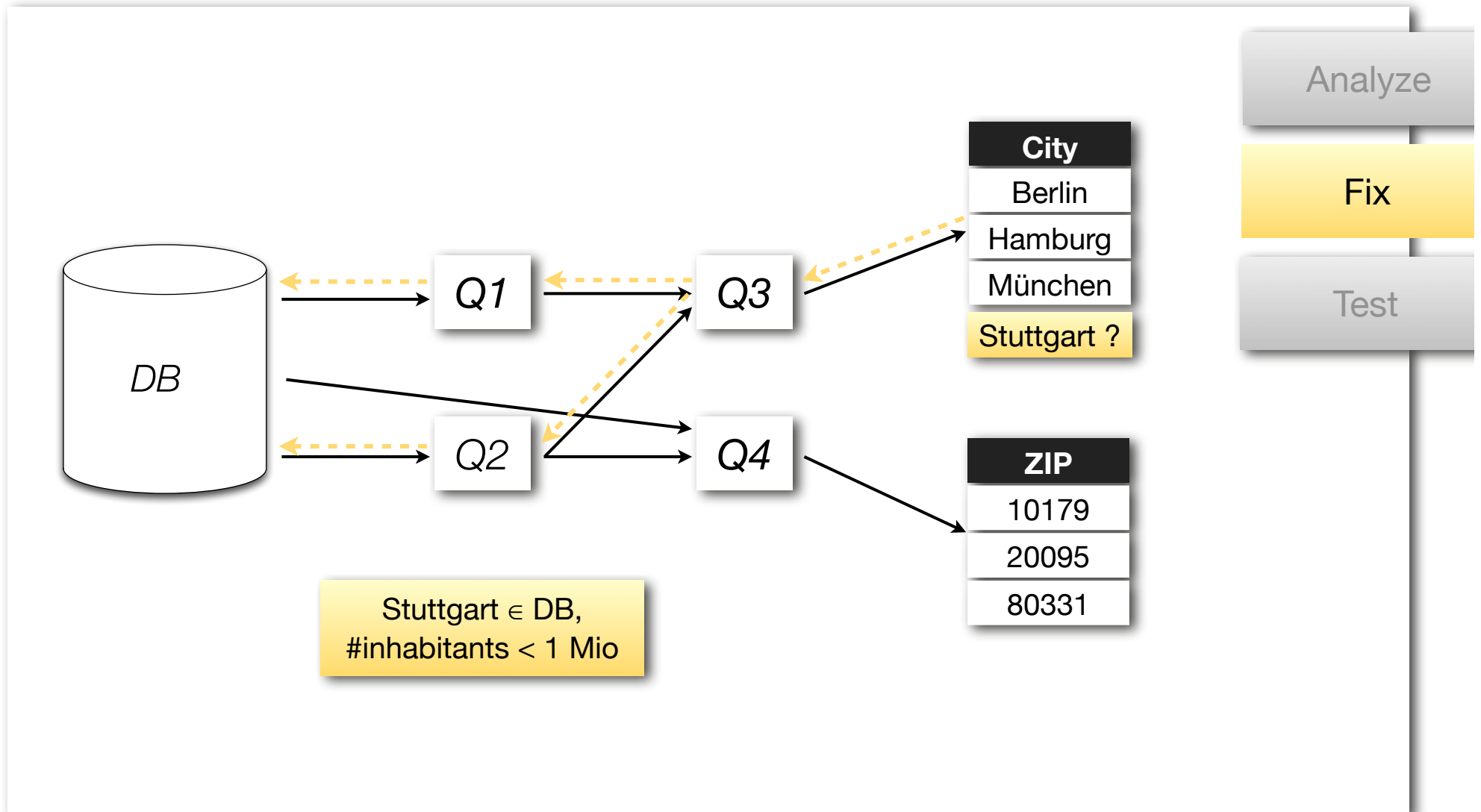
Sample Workflow



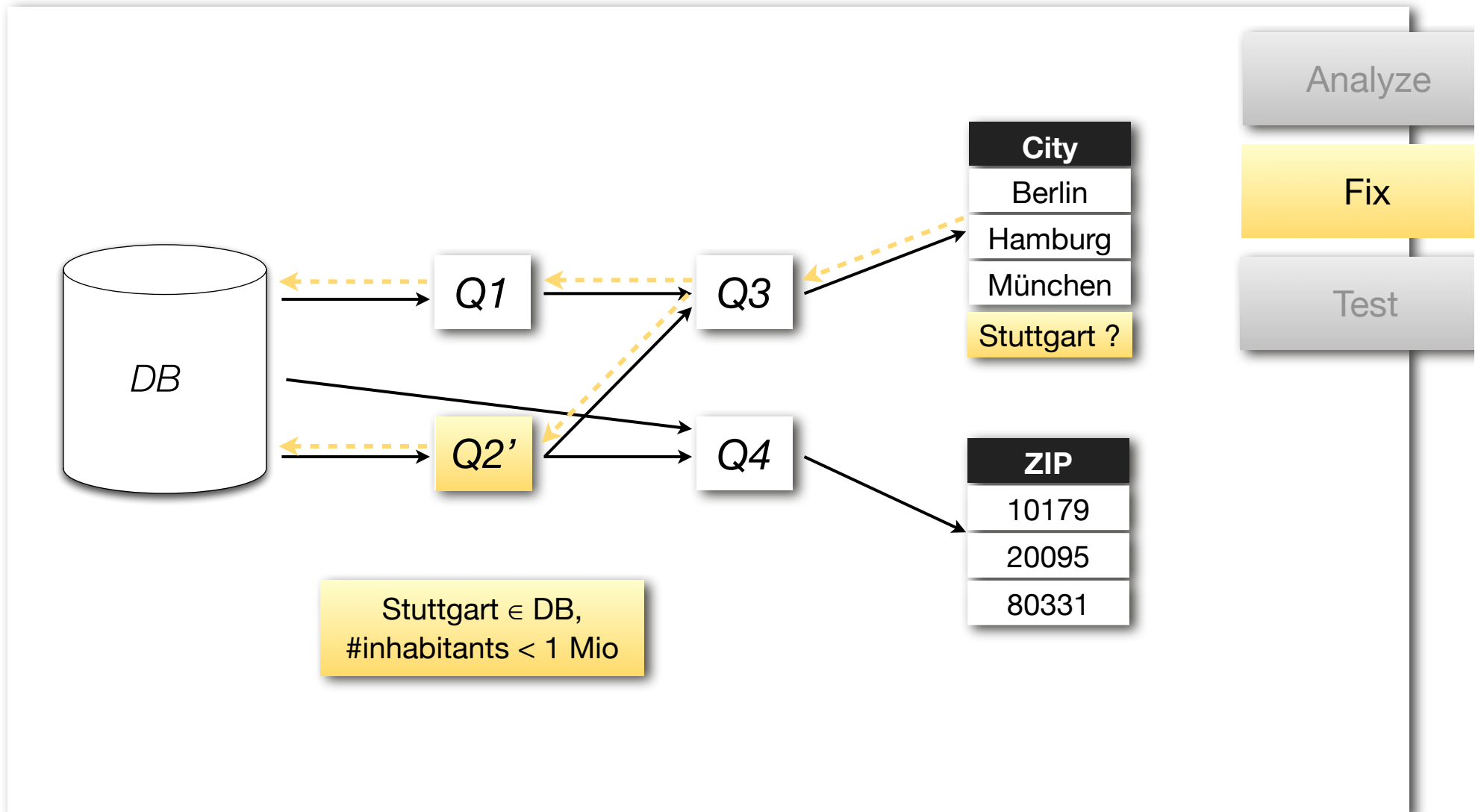
Sample Workflow



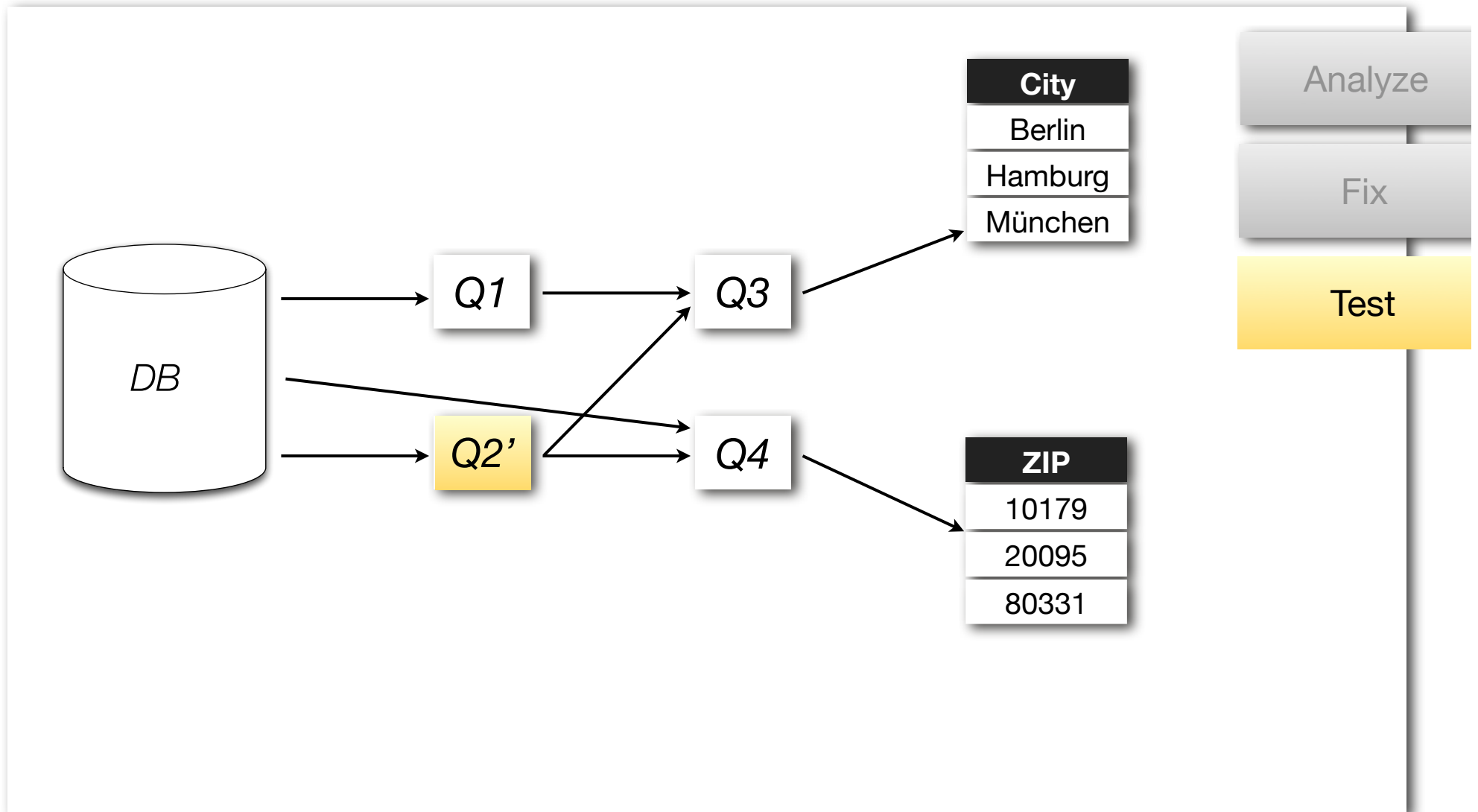
Sample Workflow



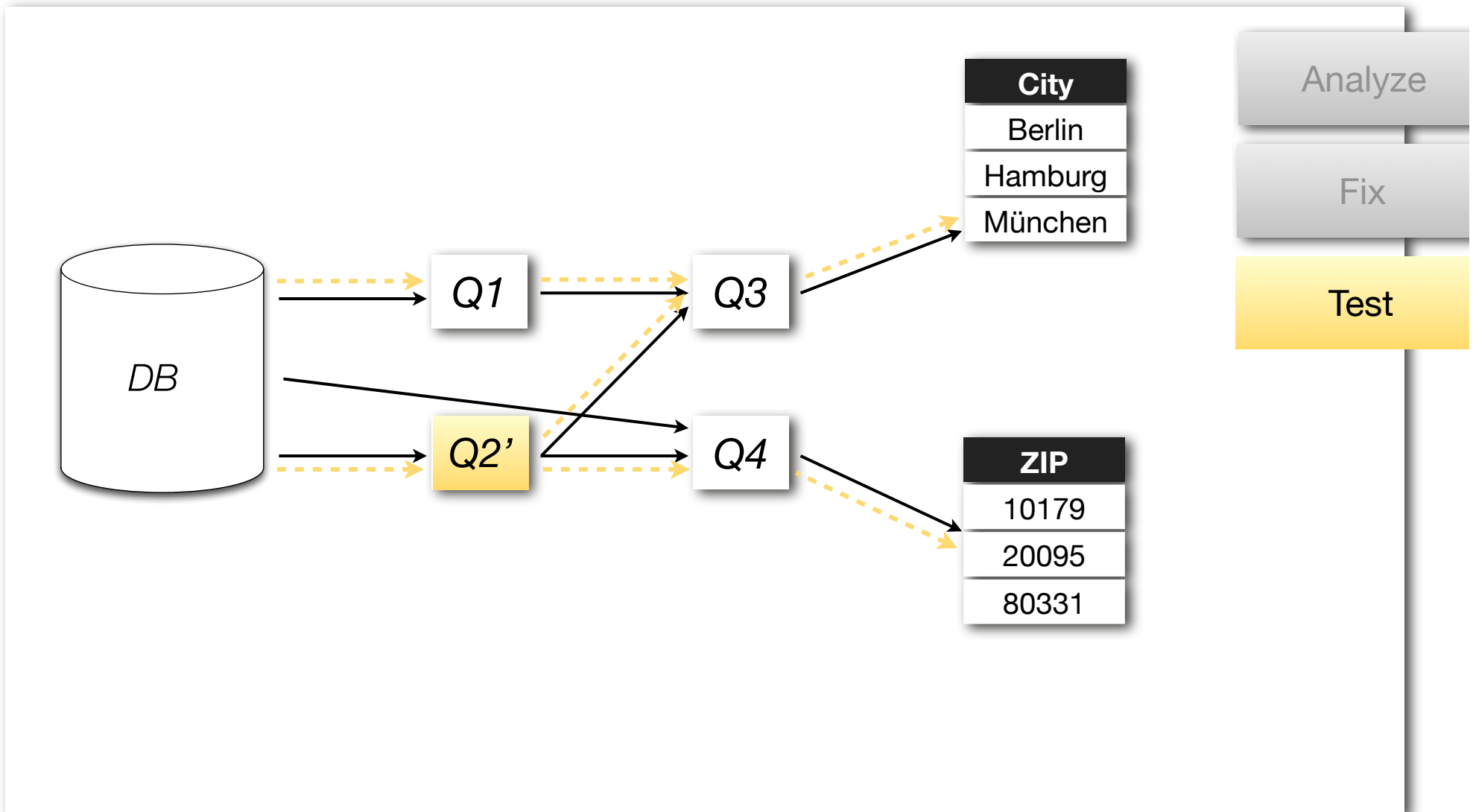
Sample Workflow



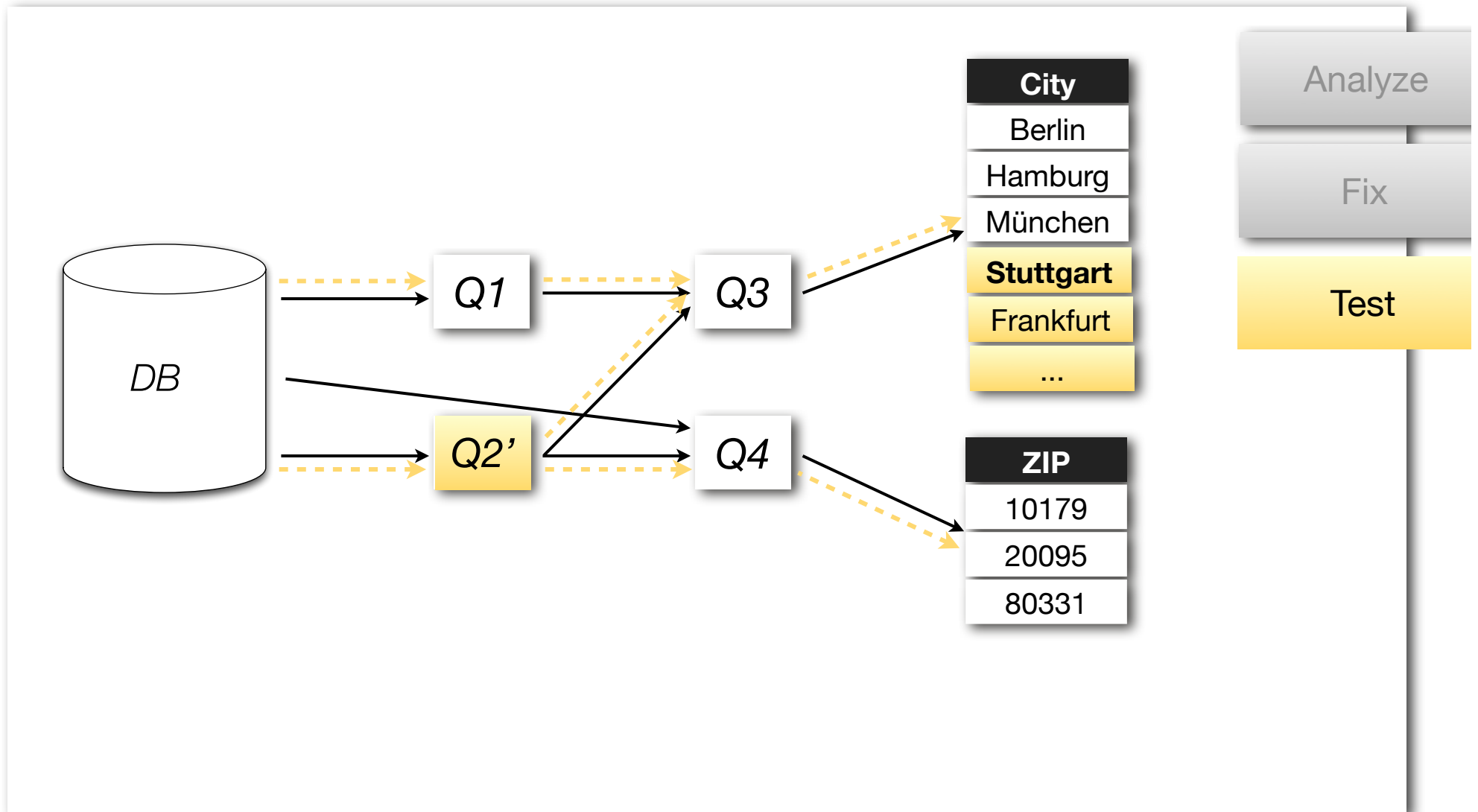
Sample Workflow



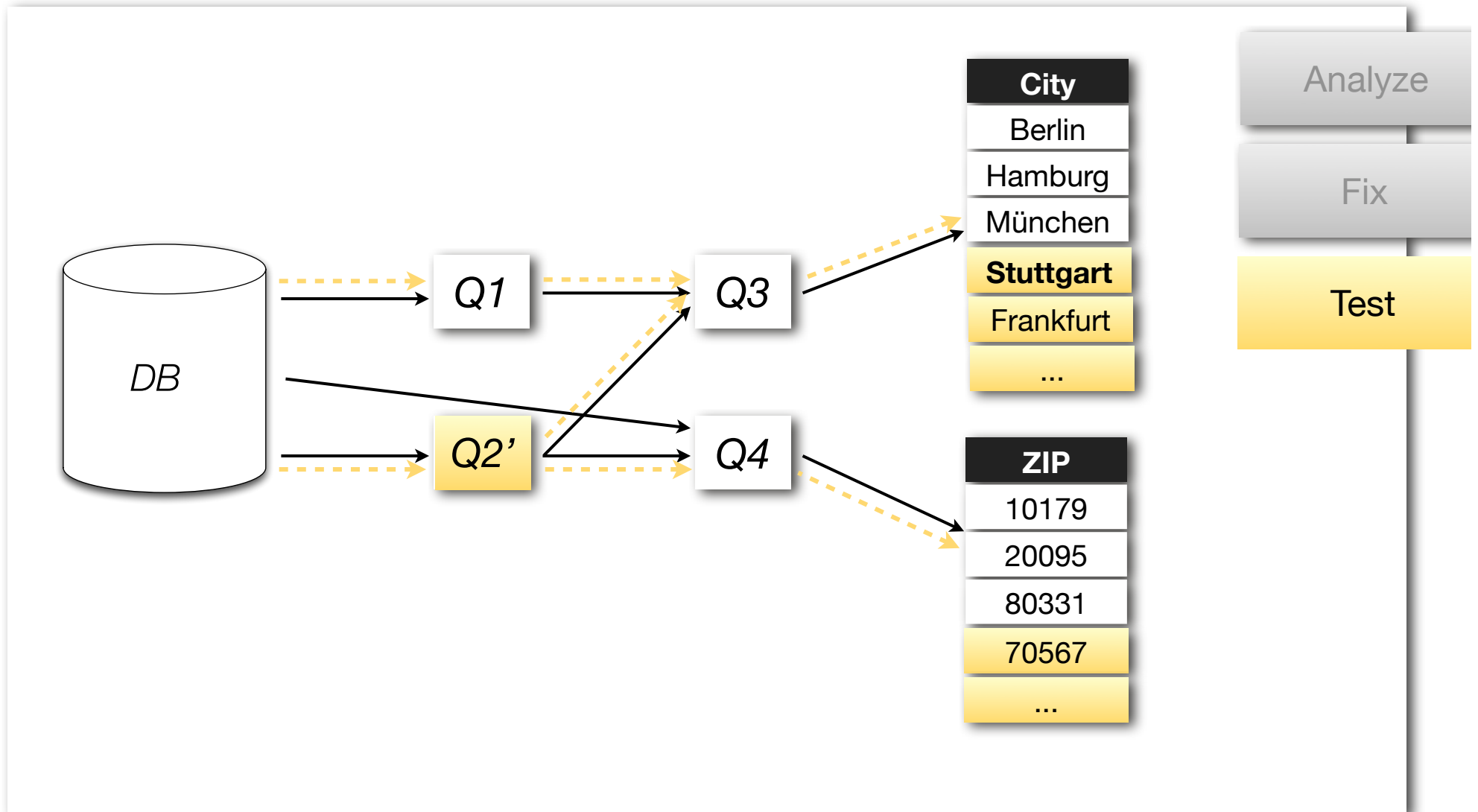
Sample Workflow



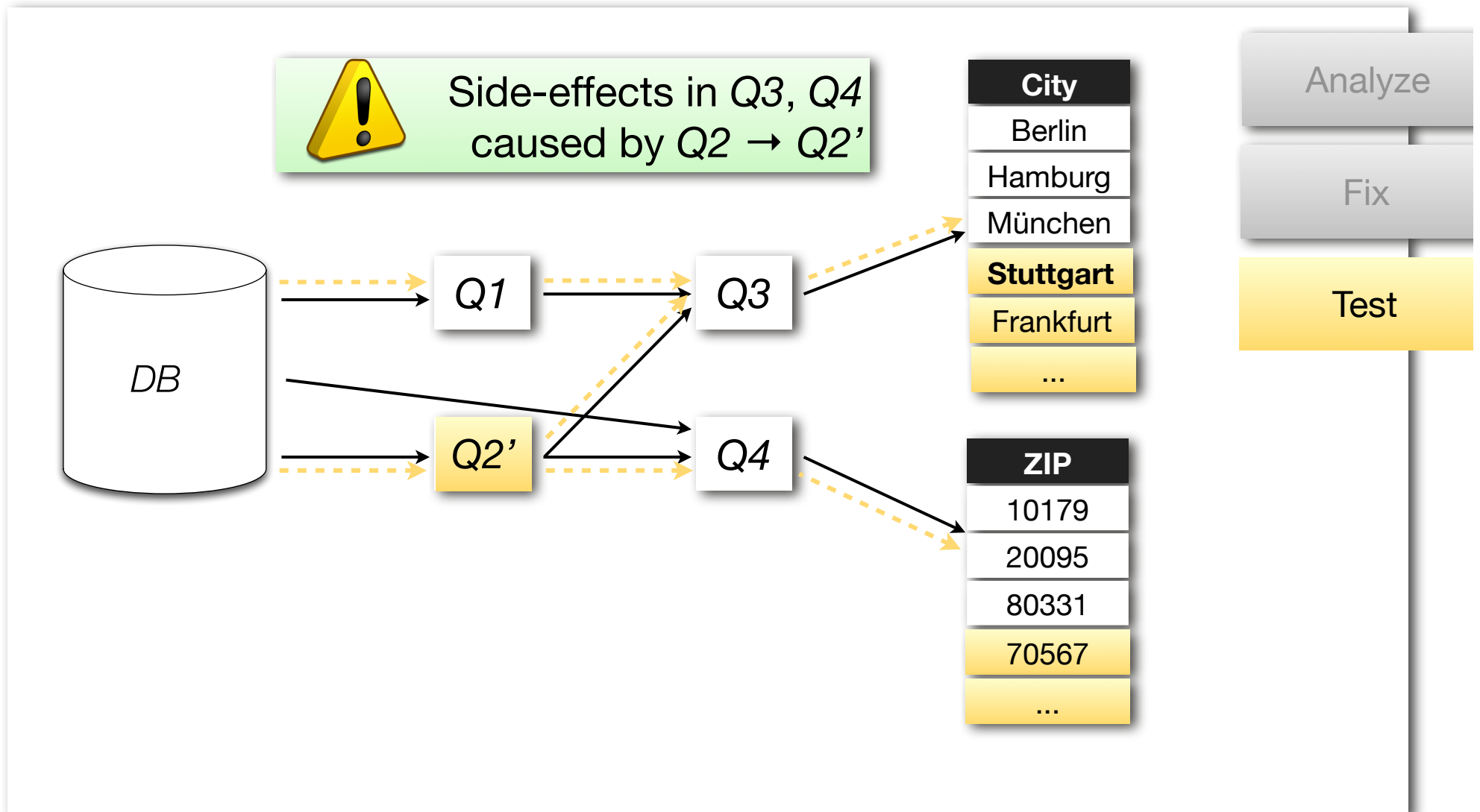
Sample Workflow



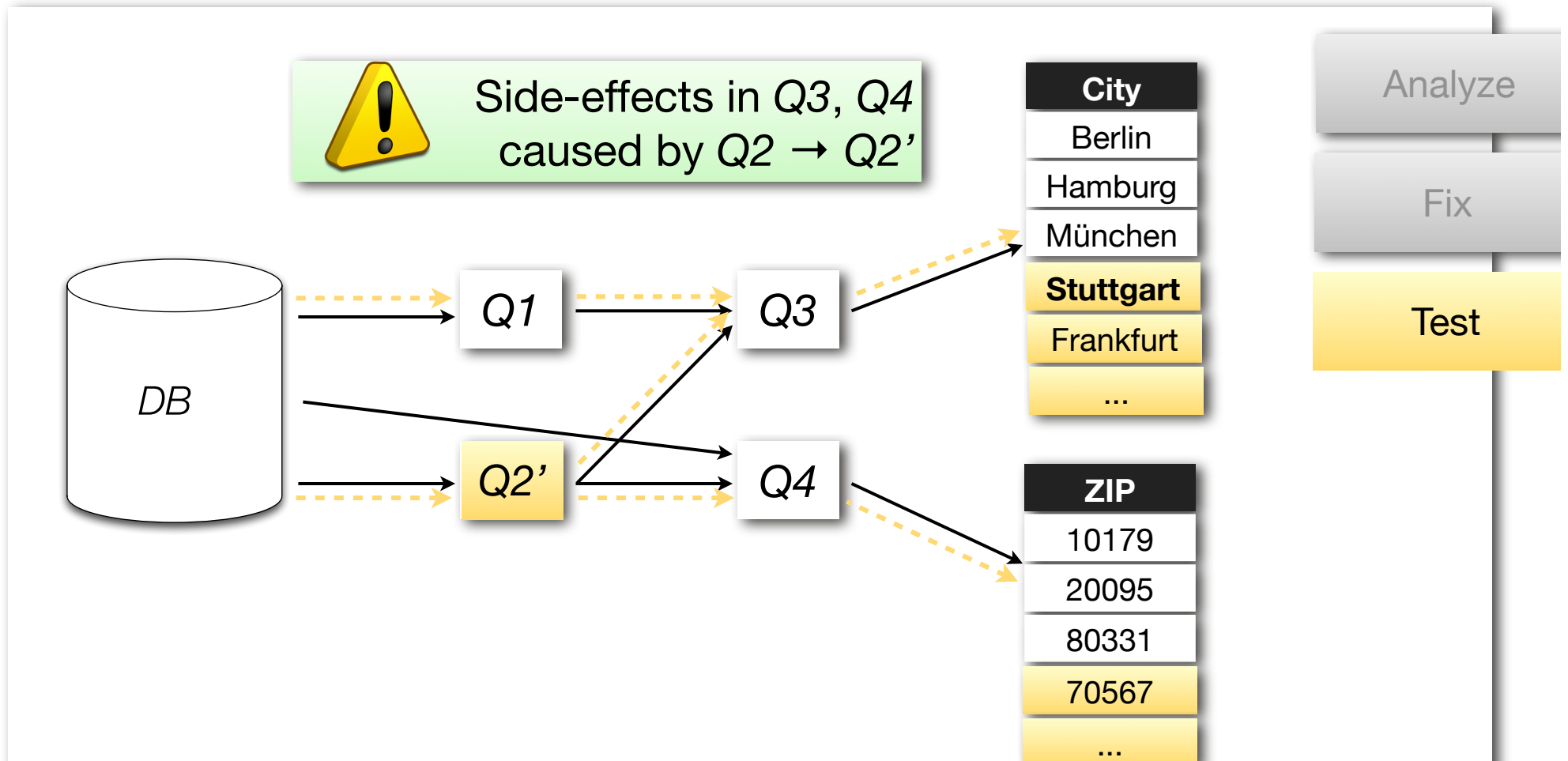
Sample Workflow



Sample Workflow



Sample Workflow



Nautilus

<http://nautilus-system.org>