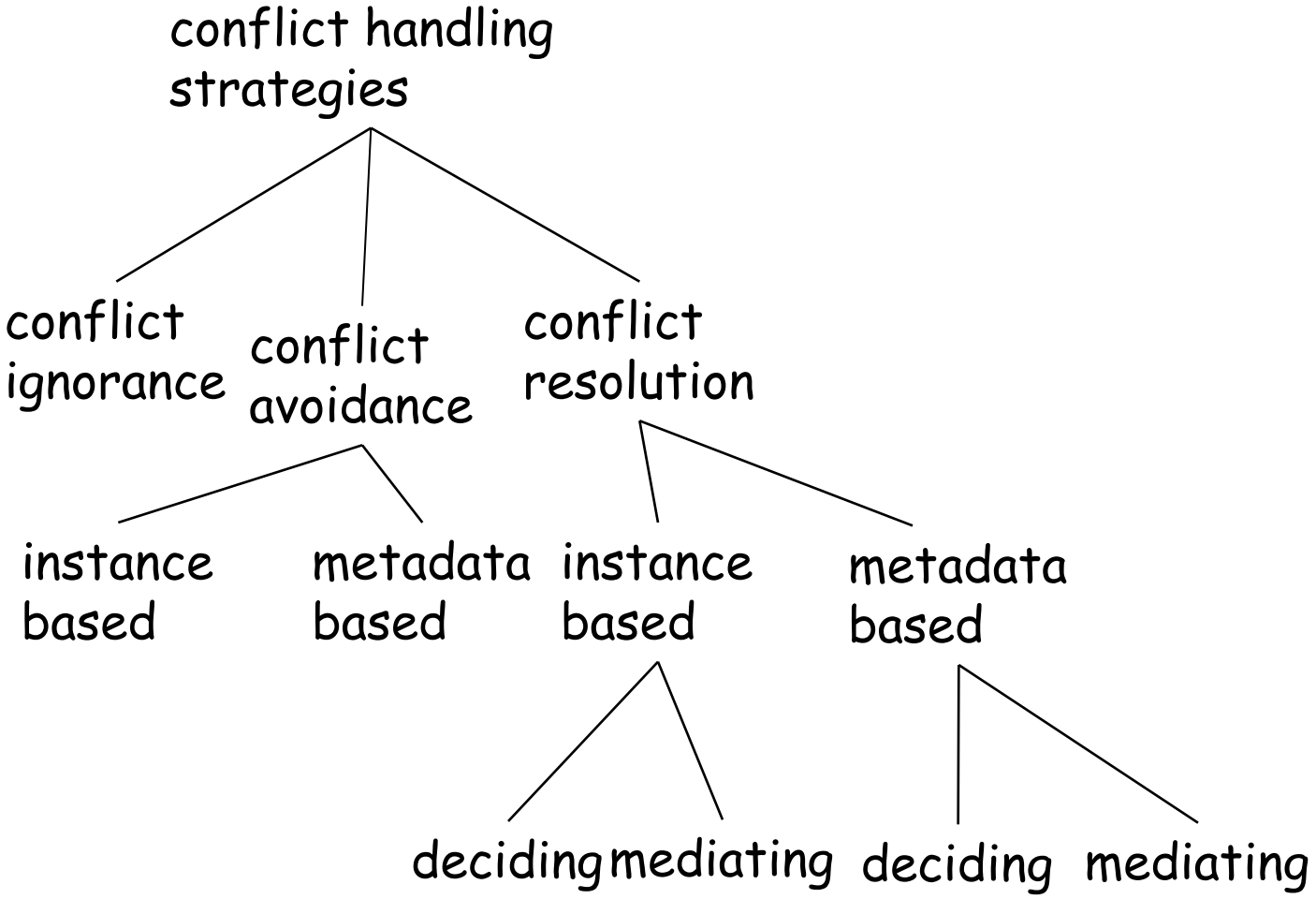
C. Batini & M. Scannapieco
Data and Information Quality Book Figures

Chapter 10: Data Quality Issues in Data Integration Systems

Data Fusion Strategies



Mediator-wrapper architecture

USER

MEDIATOR

WRAPPER

WRAPPER

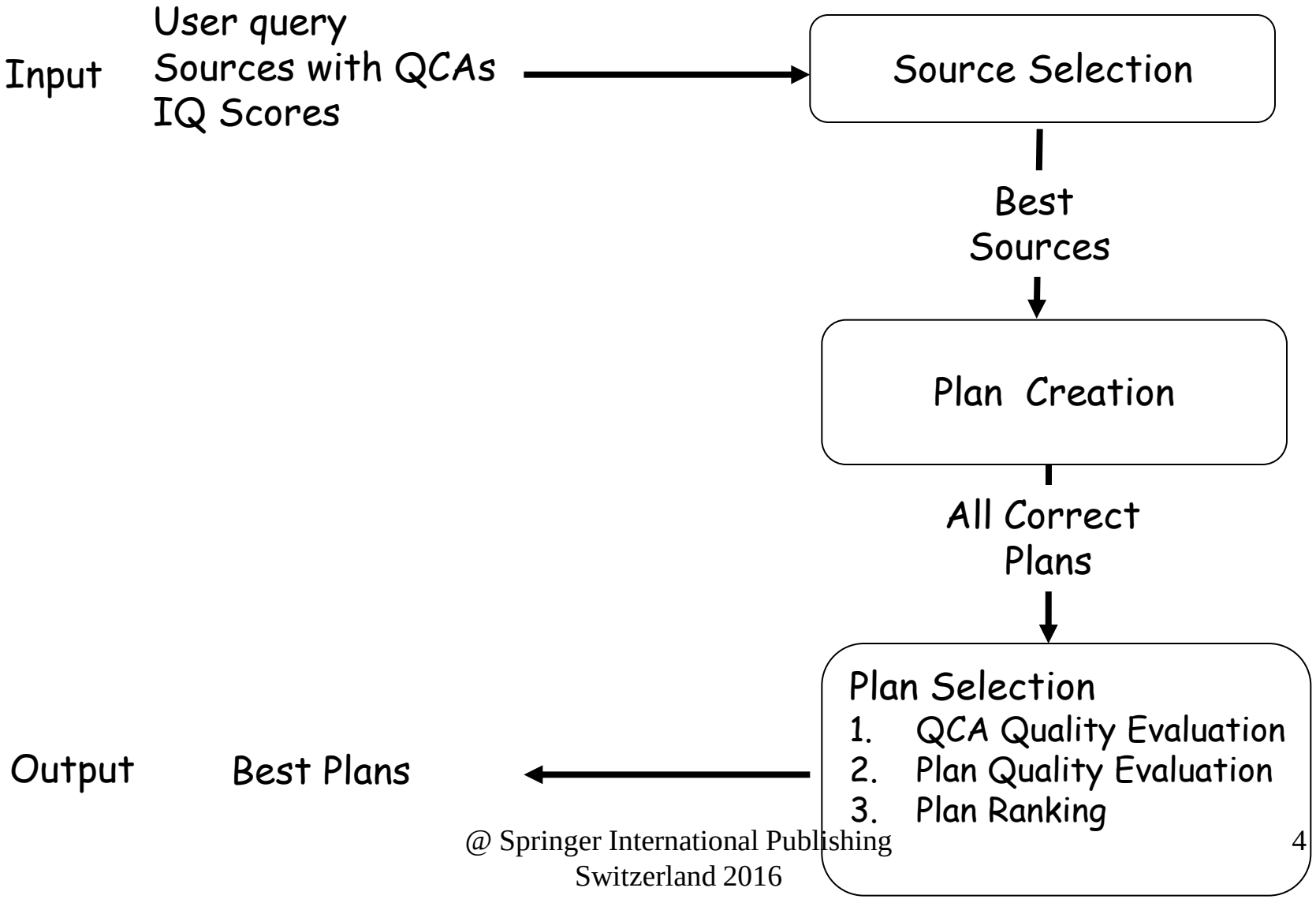
WRAPPER

DB1

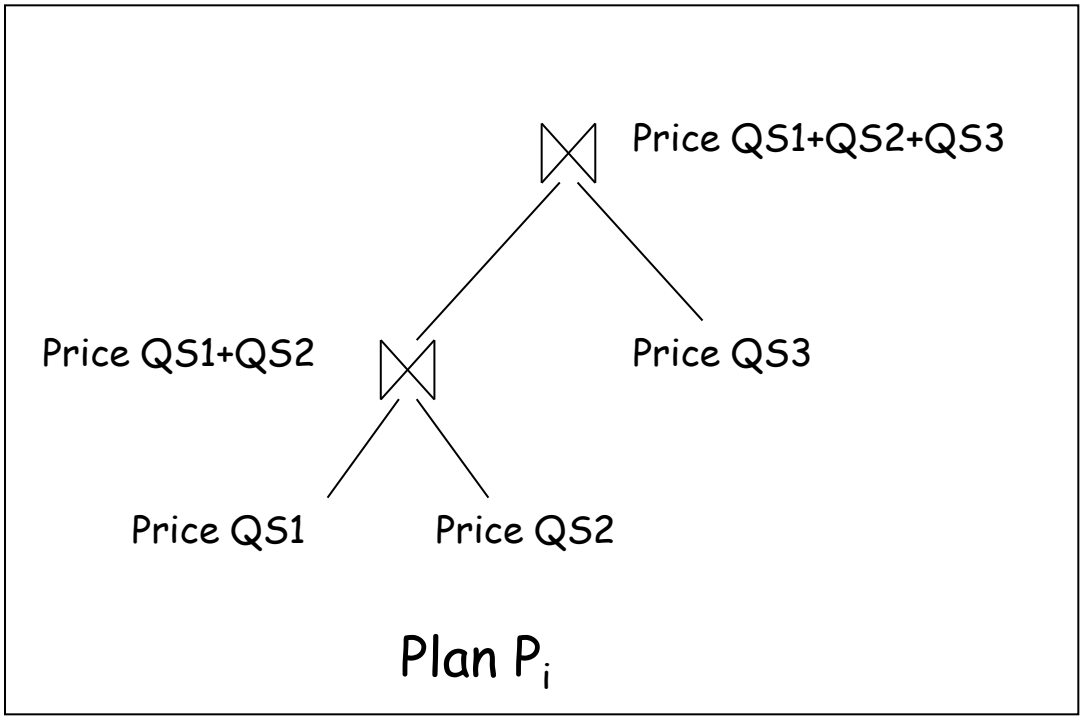
DB2

DB3

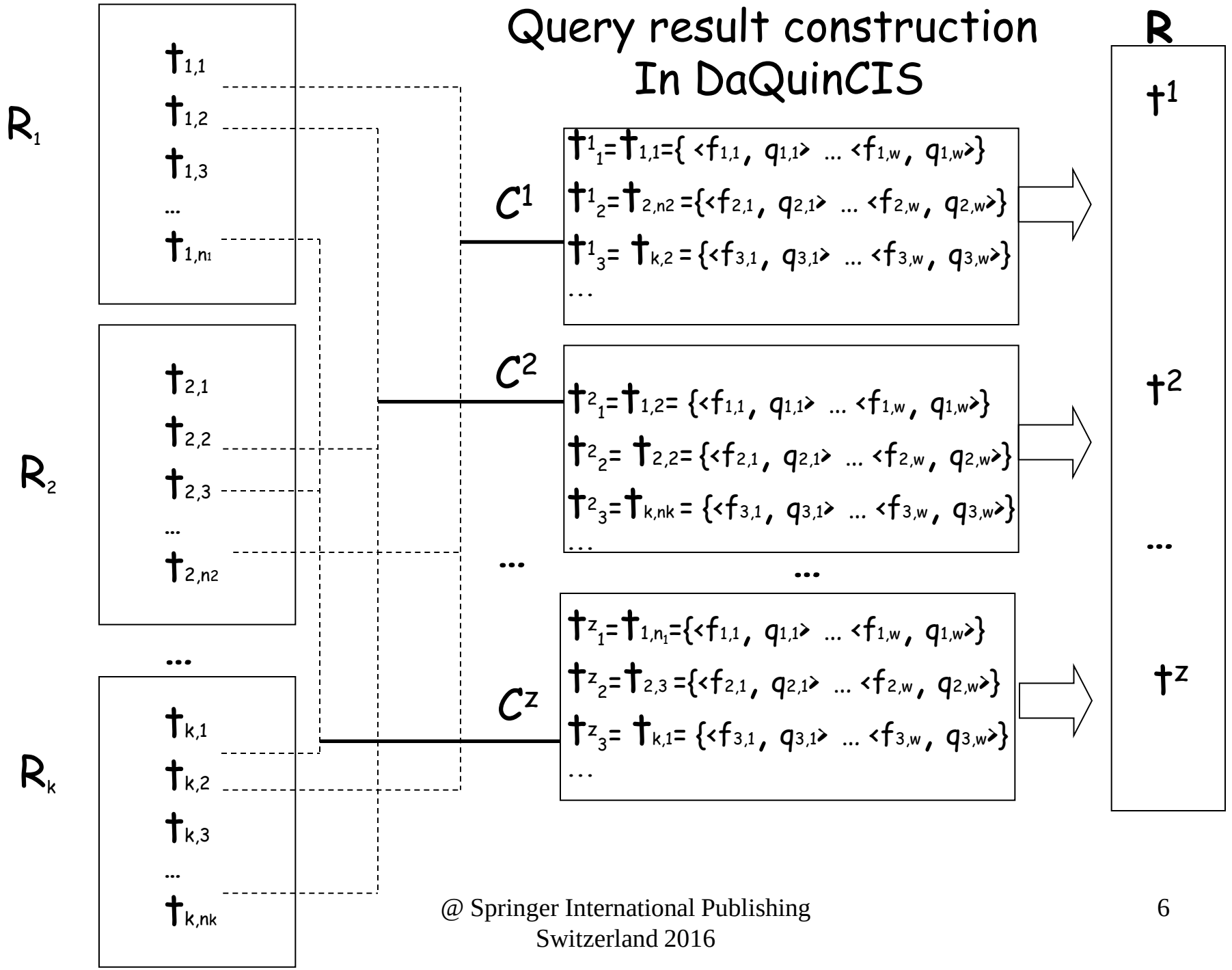
Phases of the QP-alg approach



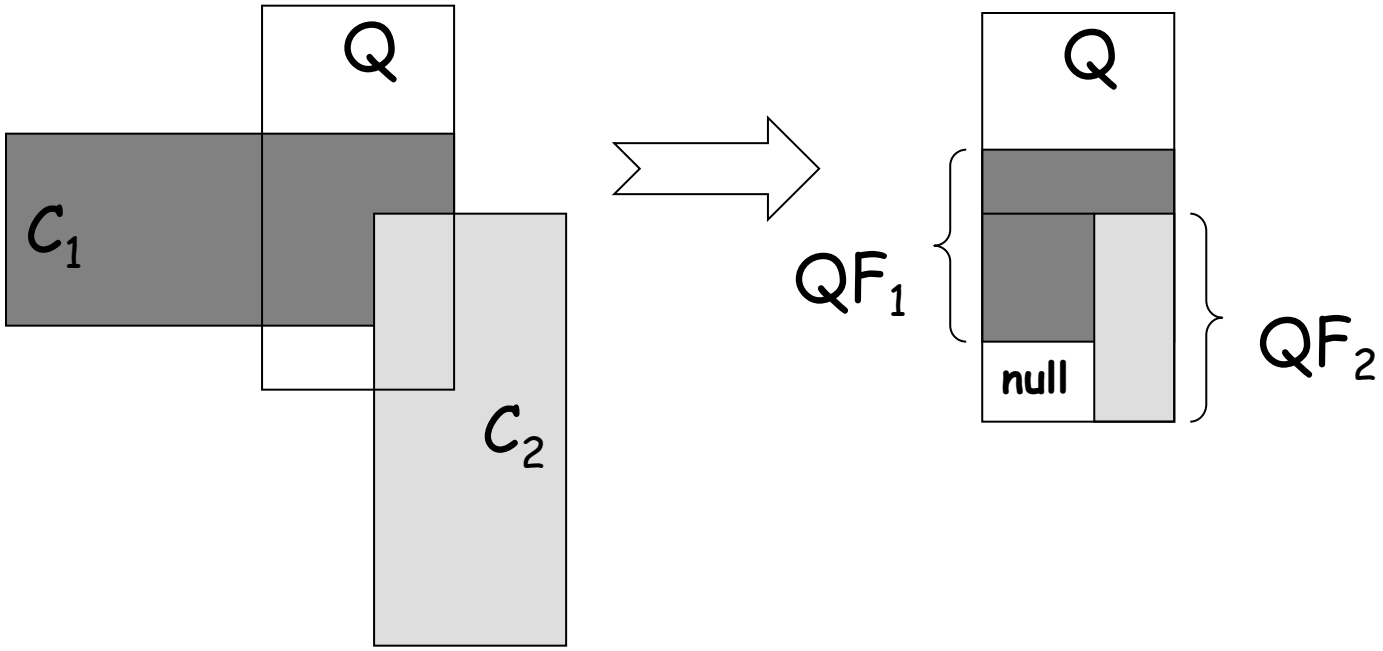
Example of Price computation for the Plan P_i



Query result construction In DaQuinCIS



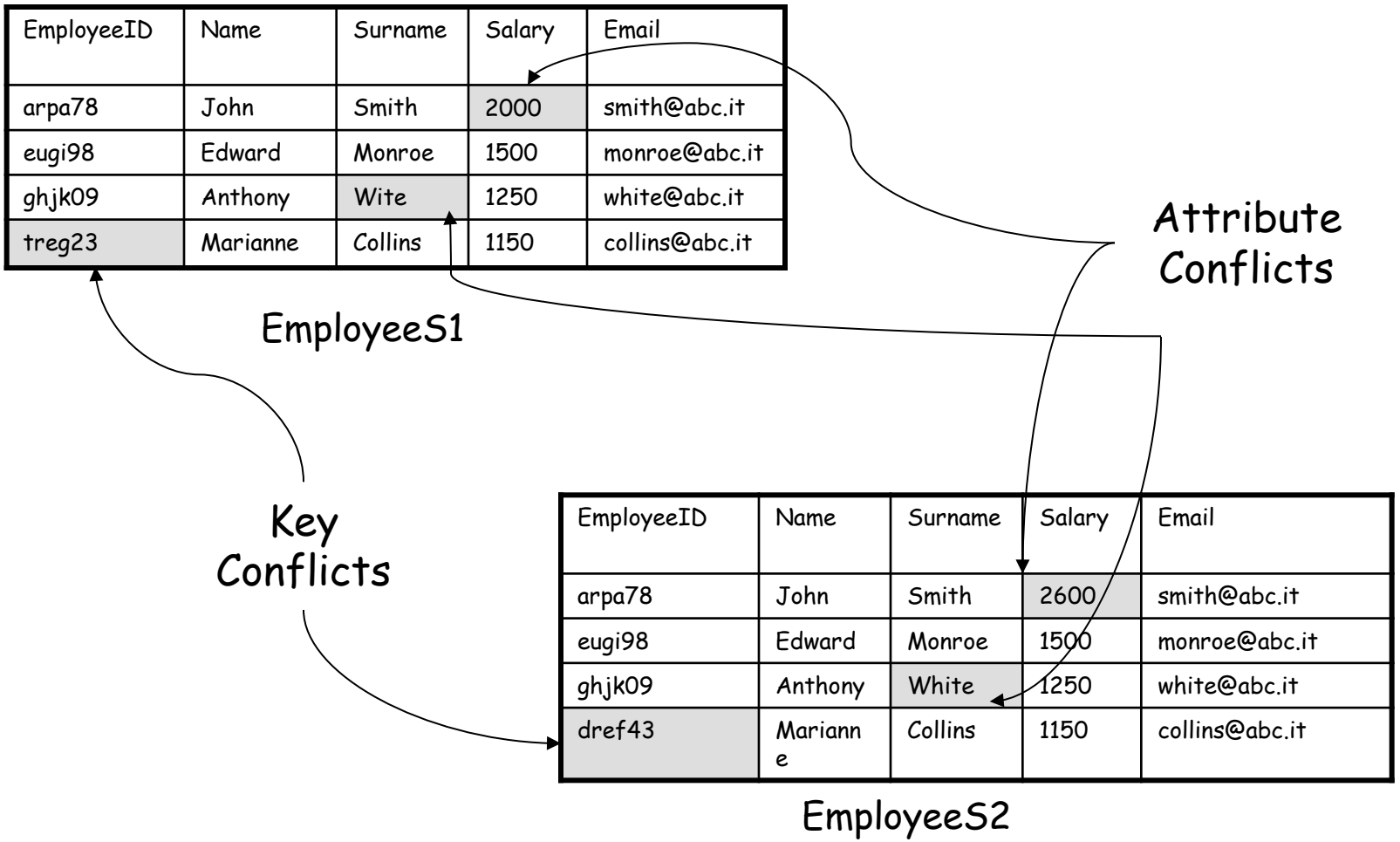
Example of query fragment construction from contributing views



Comparison of quality-driven processing techniques

Techniques	Quality Metadata	Granularity of Quality Characterization	Type of Mapping	Quality Algebra Support
QP-alg	YES	Source, Query Correspondences Assertions, User Queries	GLAV	Preliminary
DaQuinCIS Query Processing	YES	Each data element of a semistructured data model	GAV	No
FusionPlex Query Processing	YES	Source	GLAV	Preliminary

An example of key- and attribute-level conflicts



Resolution functions as proposed in [462]

Function	Attribute Type	Description
COUNT	any	Counts number of conflicting values
MIN	any	Minimum value
MAX	any	Maximum value
RANDOM	any	Random non null value
CHOOSE(Source)	any	Chooses most reliable source for the particular attribute
MAXIQ	any	Value of highest information quality
GROUP	any	Groups all conflicting values
SUM	numerical	Sums all values
MEDIAN	numerical	Median value, namely having the same number of higher and lower values
AVG	numerical	Arithmetic mean of all values
VAR	numerical	Variance of values
STDDEV	numerical	Standard Deviation of values
SHORTEST	non-numerical	Minimum length value, ignoring spaces
LONGEST	non-numerical	Maximum length value, ignoring spaces
CONCAT	non-numerical	Concatenation of values
ANNCONCAT	non-numerical	Annotated concatenation of values, whose purpose is to specify the source, before the actual returned value

Instance of the global relation Employee

TupleID	EmployeeID	Name	Surname	Salary	Email
t_1	arpa78	John	Smith	2000	smith@abc.it
t_2	eugi98	Edward	Monroe	1500	monroe@abc.it
t_3	ghjk09	Anthony	Wite	1250	white@abc.it
t_4	treg23	Marianne	Collins	1150	collins@abc.it
t_5	arpa78	John	Smith	2600	smith@abc.it
t_6	eugi98	Edward	Monroe	1500	monroe@abc.it
t_7	ghjk09	Anthony	White	1250	white@abc.it
t_8	dref43	Marianne	Collins	1150	collins@abc.it

Resolution of attribute conflicts

TupleID	EmployeeID	Name	Surname	Salary	Email
t_1	arpa78	John	Smith	2000	smith@abc.it
t_2	eugi98	Edward	Monroe	1500	monroe@abc.it
t_3	ghjk09	Anthony	White	1250	white@abc.it
t_4	treg23	Marianne	Collins	1150	collins@abc.it

RAC(Employee,Salary(MIN), Surname(Longest), EmployeeID(Any))

Resolution of tuple conflicts

TupleID	EmployeeID	Name	Surname	Salary	Email
t_1	arpa78	John	Smith	2600	smith@abc.it
t_2	eugi98	Edward	Monroe	1500	monroe@abc.it
t_3	ghjk09	Anthony	Wite	1250	white@abc.it
t_4	dref43	Marianne	Collins	1150	collins@abc.it

RTC(Employee,ANY)

Result of the context-aware query as applied to the table EmployeeS1

EmployeeID	Salary
arpa78	2000

EmployeeID	Salary
eugi98	1500

EmployeeID	Salary
ghjk09	1250
treg23	1150

Conflict resolution techniques

Techniques	Tolerance Strategies	Query Model
SQL-Based Conflict Resolution	NO	SQL
Aurora	High Confidence, RandomEvidence, PossibleAtAll	Ad-hoc Conflict Tolerant Query Model
FusionPlex	No resolution strategy, selective attribute resolution	Extended SQL
DaQuinCIS	NO	Extended XML
FraQL-based Conflict Resolution	NO	Ad-hoc FraQL
OO_{RA}	Thresholds for tolerable and intolerable conflicts	Ad hoc Object Oriented Extension (OO_{RA})