INTRODUCTION TO BUSINESS INTELLIGENCE

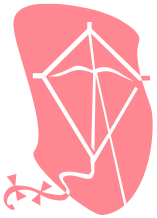
Lecture 6

Agenda

- Rules for slowly changing dimensions
- Data integration methods
- Data cleansing

Slowly changing dimensions

Types of slowly changing dimensions



- SCD - Slowly Changing Dimensions.
- They make it possible to manage methodological or other changes that occur in a time series.
- Applies to terms contained in dimension tables.

- The most common SCD:
  - Type 1 - attribute value override
  - Type 2 - adding a new dimension row
  - Type 3 - adding a new dimension attribute

SCD – type 1

Means that the previous value of the attribute will be overwritten with its new value.

The historical value is lost with no possibilities to revert.

This is one of the not expected solutions as it can lead to time series disturbances if the methodology of the concept changes.

SCD – type 2

It is the most commonly used method of managing slowly changing dimensions rules.

This method creates a new attribute with its own primary key, and any previous values methodologically similar to the new attribute are kept with the primary key unchanged in the database.

SCD – type 3

It is one of the solutions that allow you to store the previous and current definition of an attribute (e.g. a changed area of a voivodeship) in one place.

Thus, the user will keep track of the time series using two definitions for a single row.

Retrospection and SCD

- true - the object will faithfully reproduce the past,
- false - as the value of the object changes, the data related to its history will also change,
- persistent - means that the value of the object does not change over time.

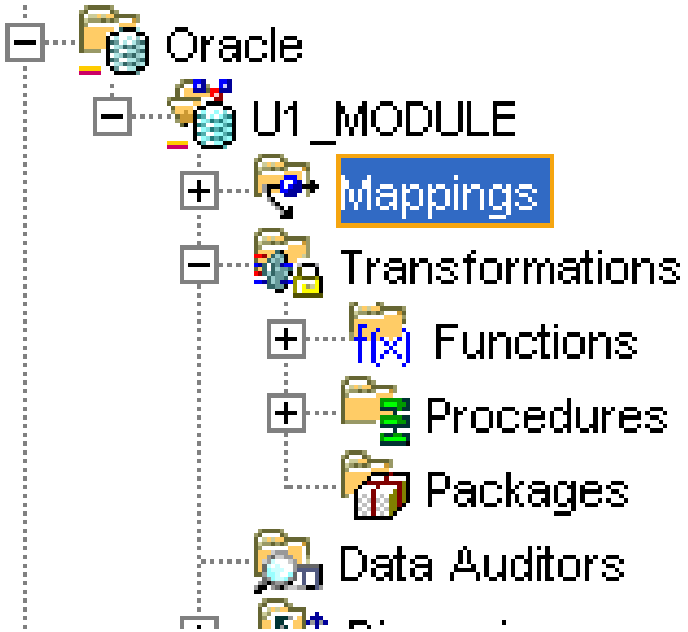
Data integration methods

Schema integration

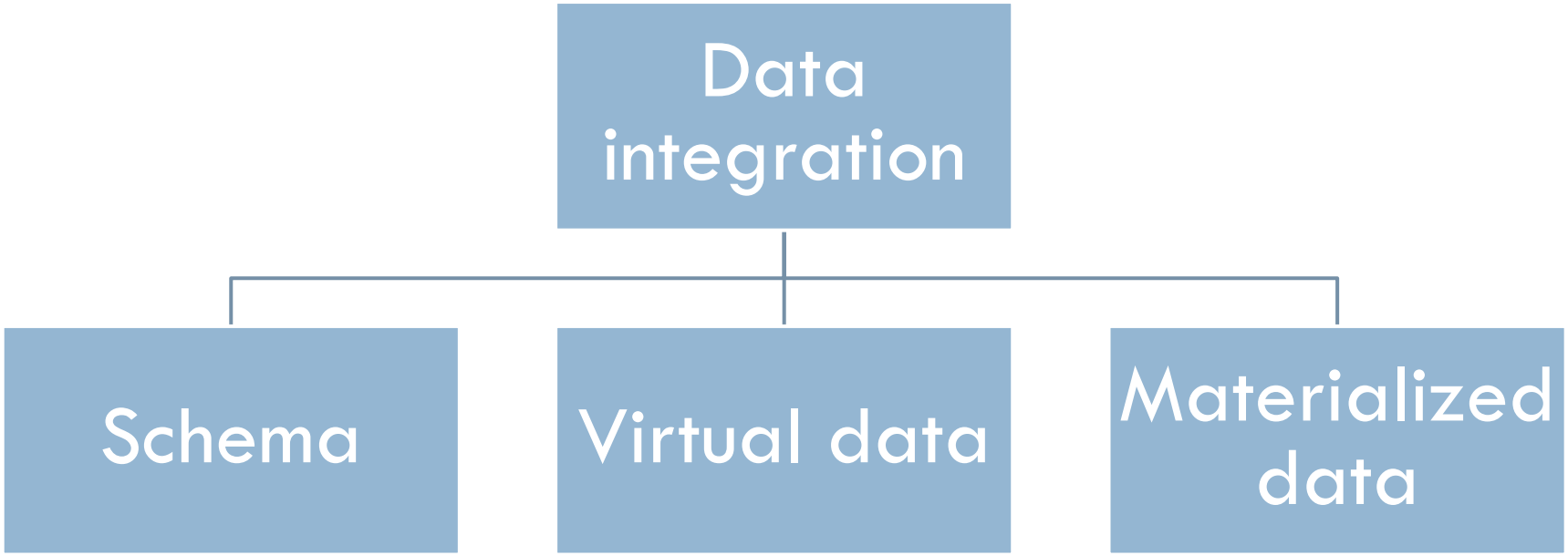
Virtual data integration

Materialized data integration

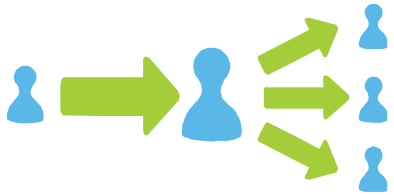
Data integration



Types of integration - approaches



Schema integration

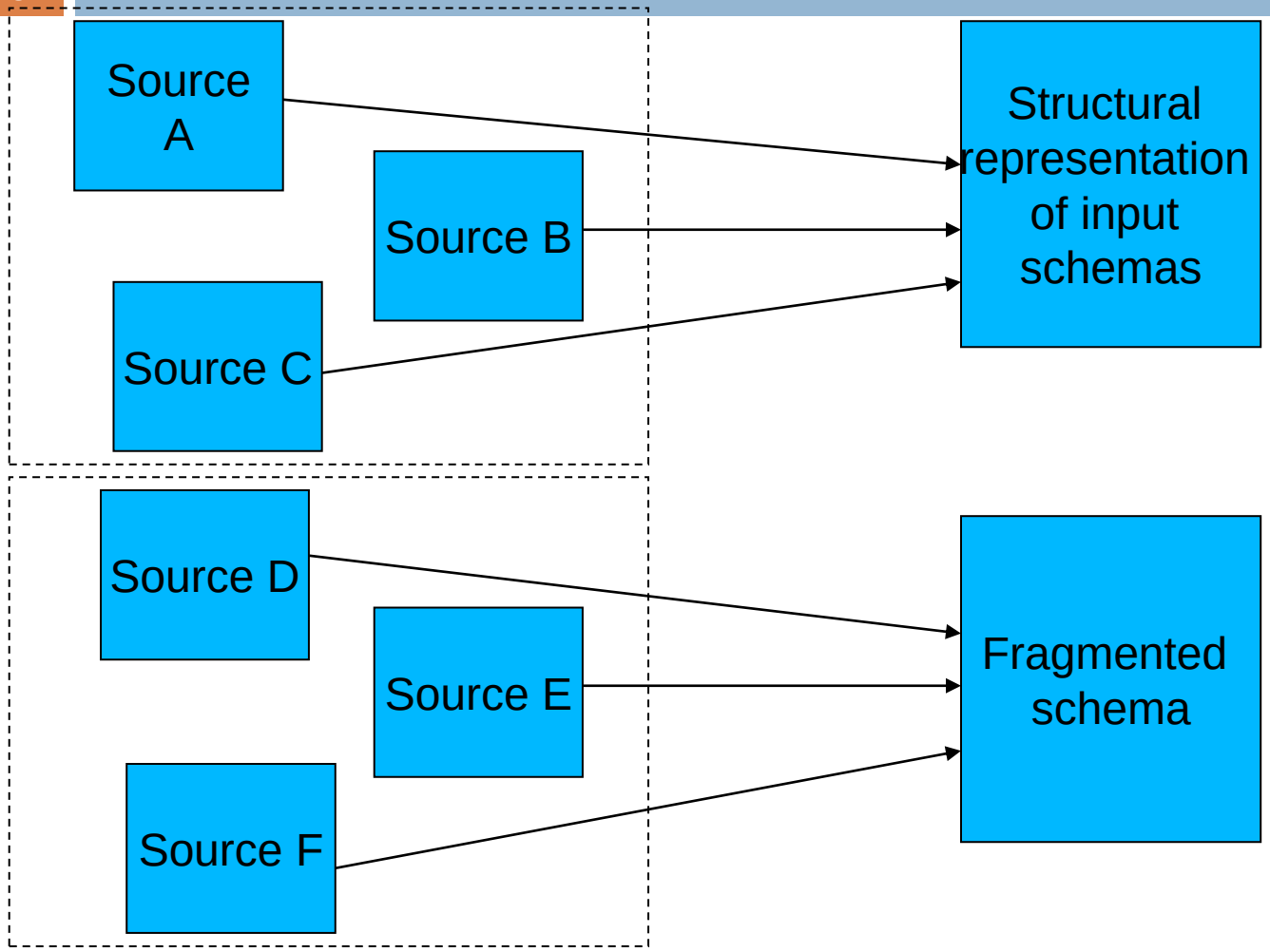


The input to the integration process is a set of source schemas.

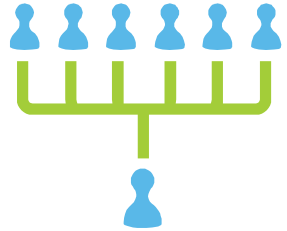
The result of the integration process - a single (target) schema, representing a uniform, structural representation of the input schemas.

The integration process also results in specifying the mapping of source schemas to fragments of the target schema.

Schema integration



Schema integration

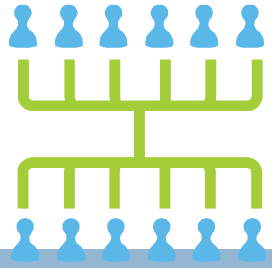


Schema integration is the process of reconciling data schemas from different sources to create a unified description of the data of interest.

Most often, this process was performed once and was aimed at creating a global schema that uniformly describes the data.

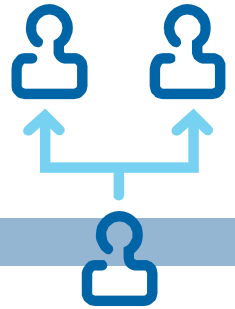
Recently, an incremental model has been used more and more often, which is of particular importance in related applications with autonomous and dynamic source databases.

Schema integration



This model is related with building a set of independent partial schemas and formalizing relationships between entities from individual schemas using so-called cross-schema assertions.

Schema integration

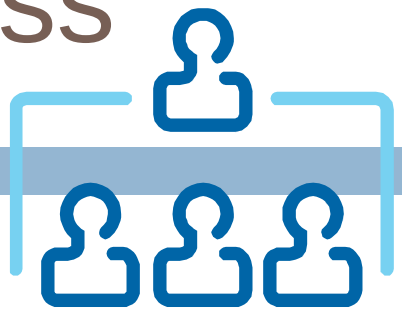


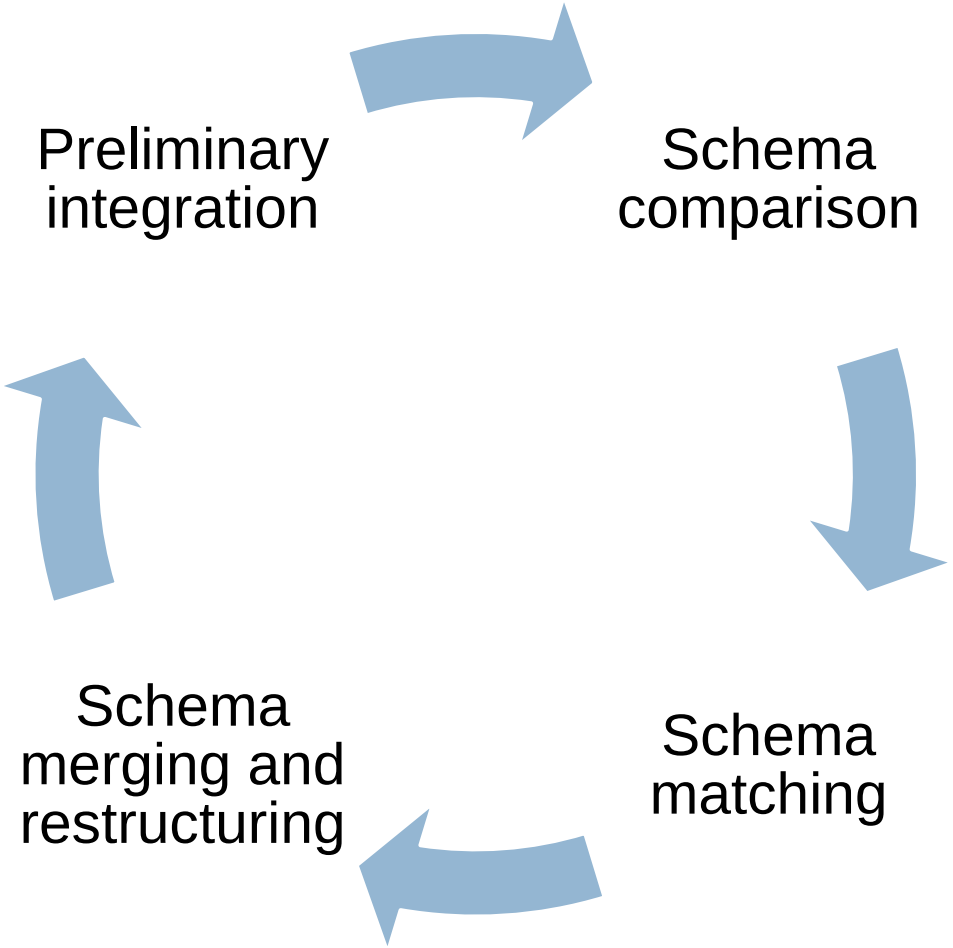
Assuming that the schemas of different data sources remain the same, the final result of the incremental approach should be the same global schema as obtained through one-time integration.

In practice such a scheme is never created, which results from the variability of sources.

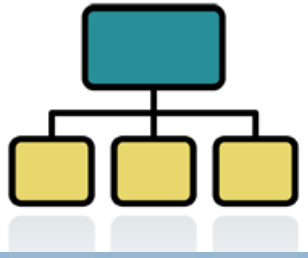
It may also happen that the integration process is partial and only some aspects or components of the source databases will be considered.

Schema integration process





Schema integration process – preliminary integration



Schema analysis to develop an overall integration strategy, including selection of schemas for integration and sequencing of integration.

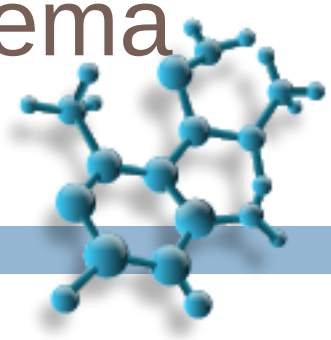
Additionally, preferences of individual schemas or their fragments can be set. This affects the later usability and adequacy of the data represented by the global schema.

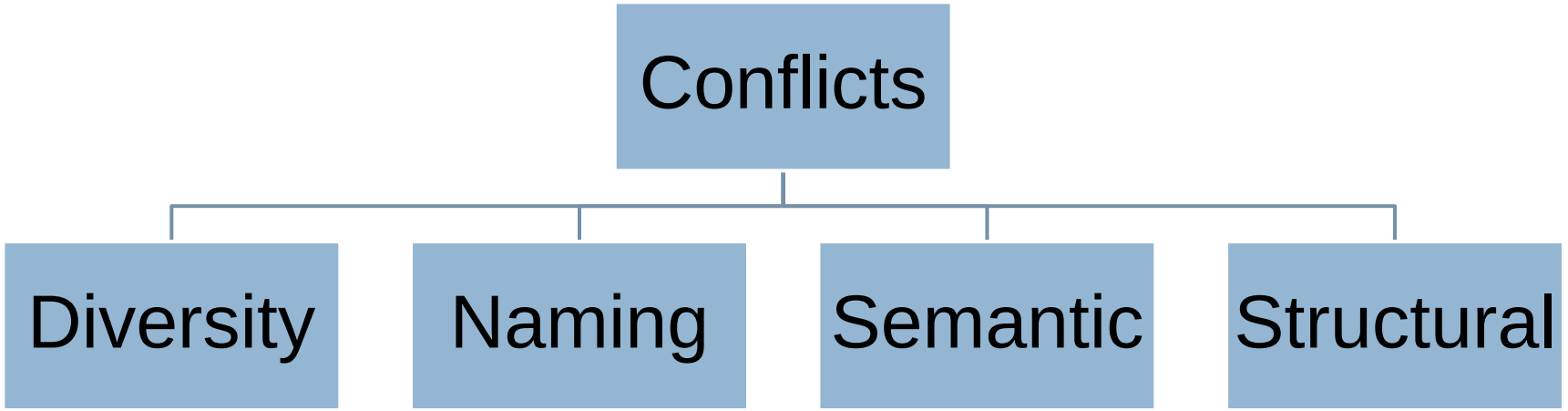
Schema integration process - schema comparison

Analysis of relationships between concepts derived from different schemas and detecting potential conflicts.

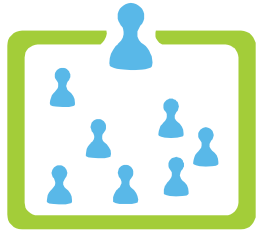
Typically, when comparing schemas, cross-schema properties are detected.

Schema integration process - schema comparison, conflicts



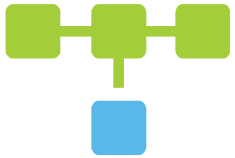


Schema integration process - schema comparison



- **diversity** conflicts - source schemas use different data models,
- **naming** conflicts - different schemas use different terminology for the same data (homonyms and synonyms of terms),
- **semantic** conflicts - similar concepts from the real world are modeled at different levels of abstraction,
- **structural** conflicts - the same concepts are represented by different structures.

Schema integration process - schema matching



Most often it takes place in a semi-automatic manner, where the conflicts reported by the system are resolved by the designer.

Schemas are often modeled as abstract data types, and the transformations of schemas are expressed in terms of the interpretation of the signature.

Schema integration process - merging and restructuring of schemas

A global schema is created as a result of overlapping matching schemas.

The schema merge technique can be performed by the binary schema merge operator expressed in the general data model.

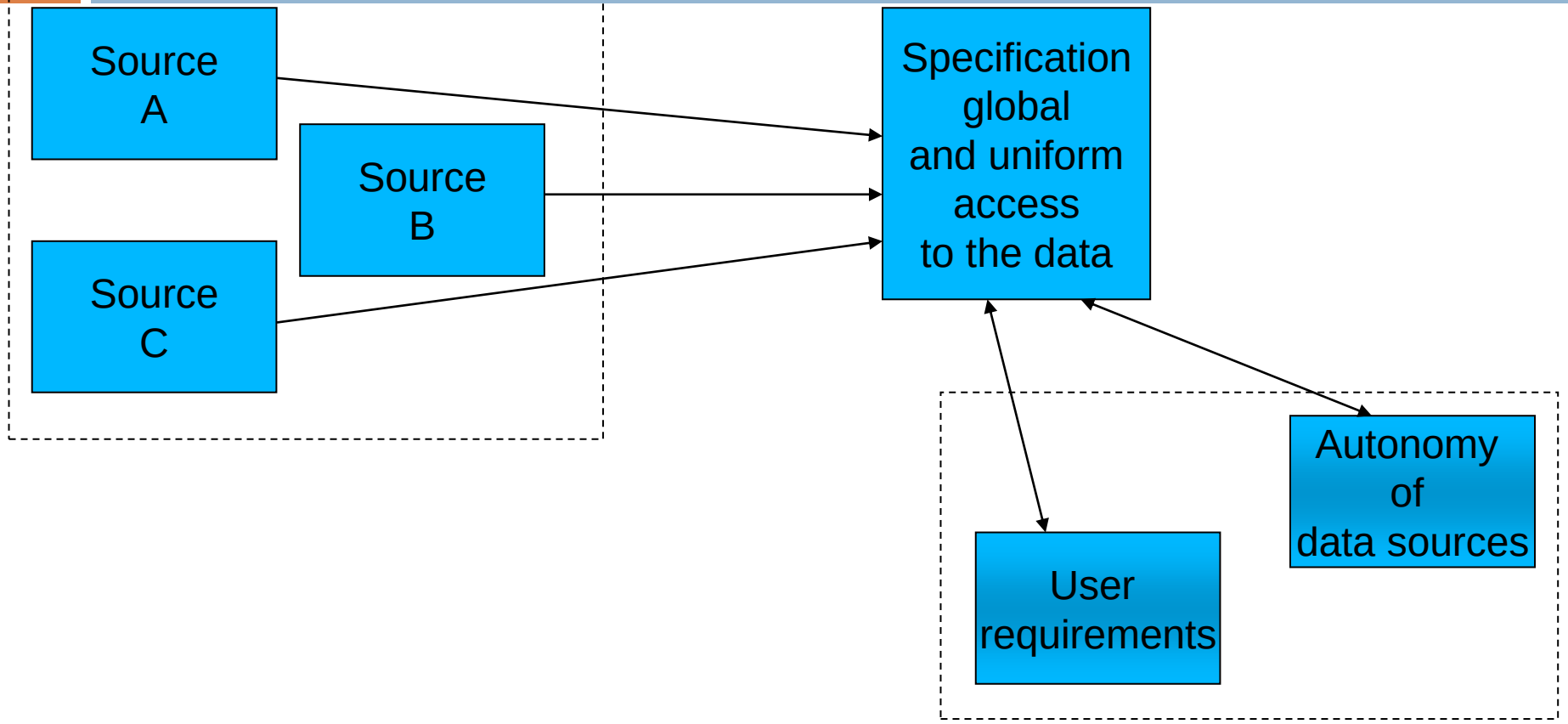
The technique called structural integration allows for the integration of objects showing structural similarities.

Virtual data integration

The input information includes the source data sets. Virtual is related to create views of these sources.

The result is a global specification and uniform access to this data, taking into account the specific needs of users and the autonomy of data sources.

Virtual data integration

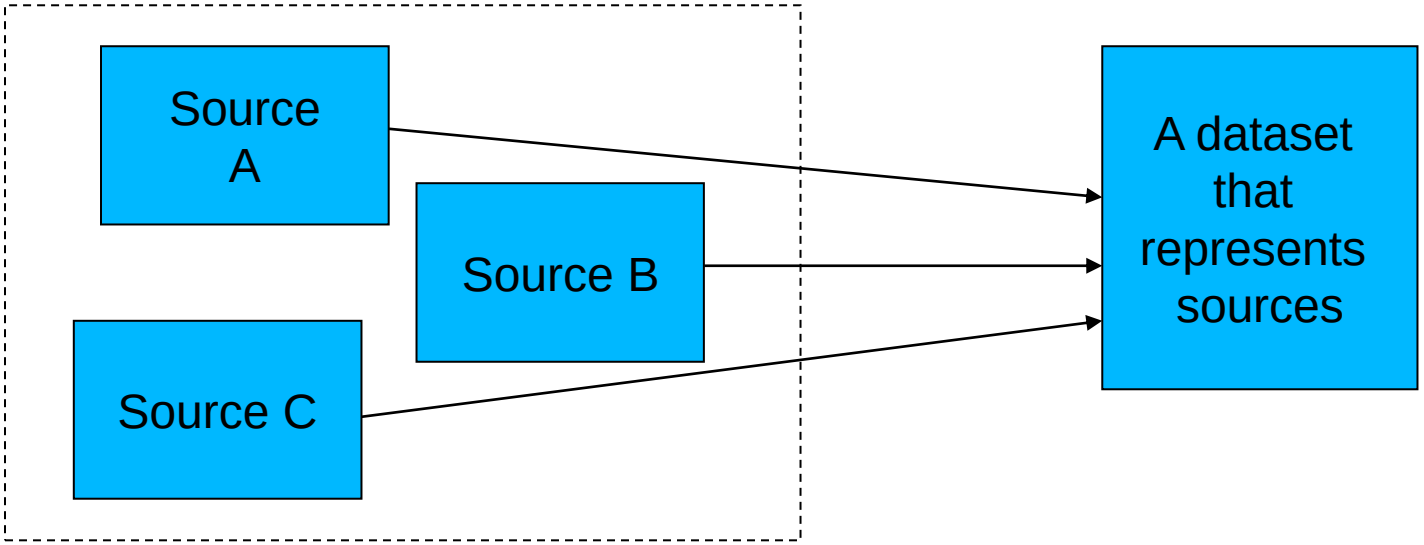


Materialized data integration

The input information includes the source data sets.

The result is a structured data set and content representation of sources.

Materialized data integration



Materialized data integration

The main focus is on handling perspectives related to updating source information.

There are so-called self-service perspectives that allow direct updating from only the data source log.

Materialized data integration

Materialized data integration can also lead to query creation where some data may be missing.

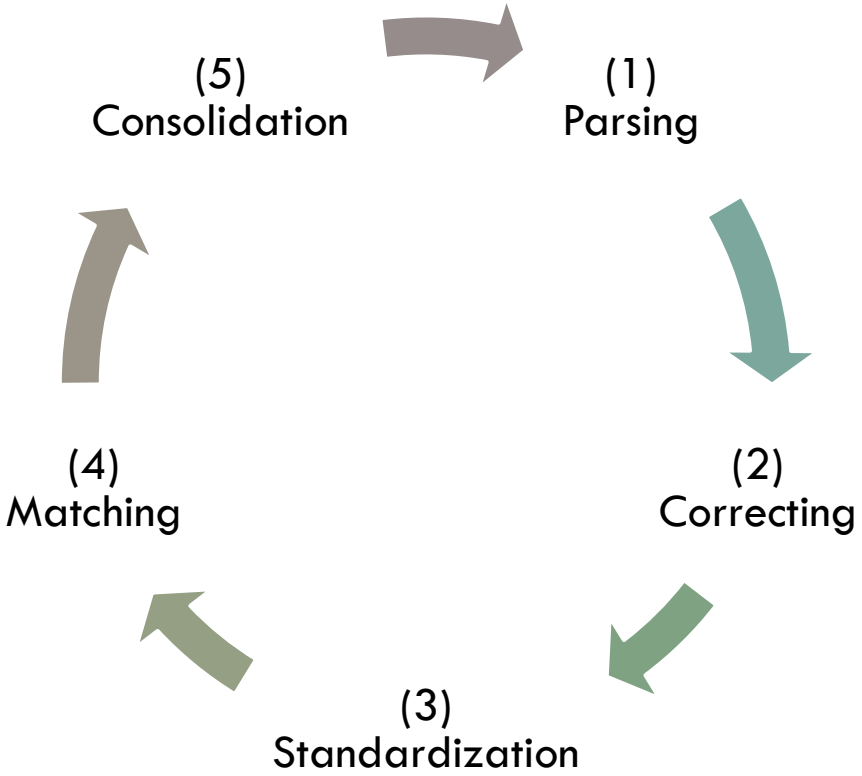
If the query forces data to be retrieved using materialized views and the data is not available, then special semantic algorithms can find this data referring to the relation not included in the materialized view but resulting from the data warehouse structure.

Data cleansing

What is dirty data?

- Wrong values
- No data
- Fields with multiple meanings
- Encrypted data
- Inconsistent data
- Incorrect address details
- Violation of business rules
- Multiple use of master keys
- Ambiguous identifiers
- Data integration problems

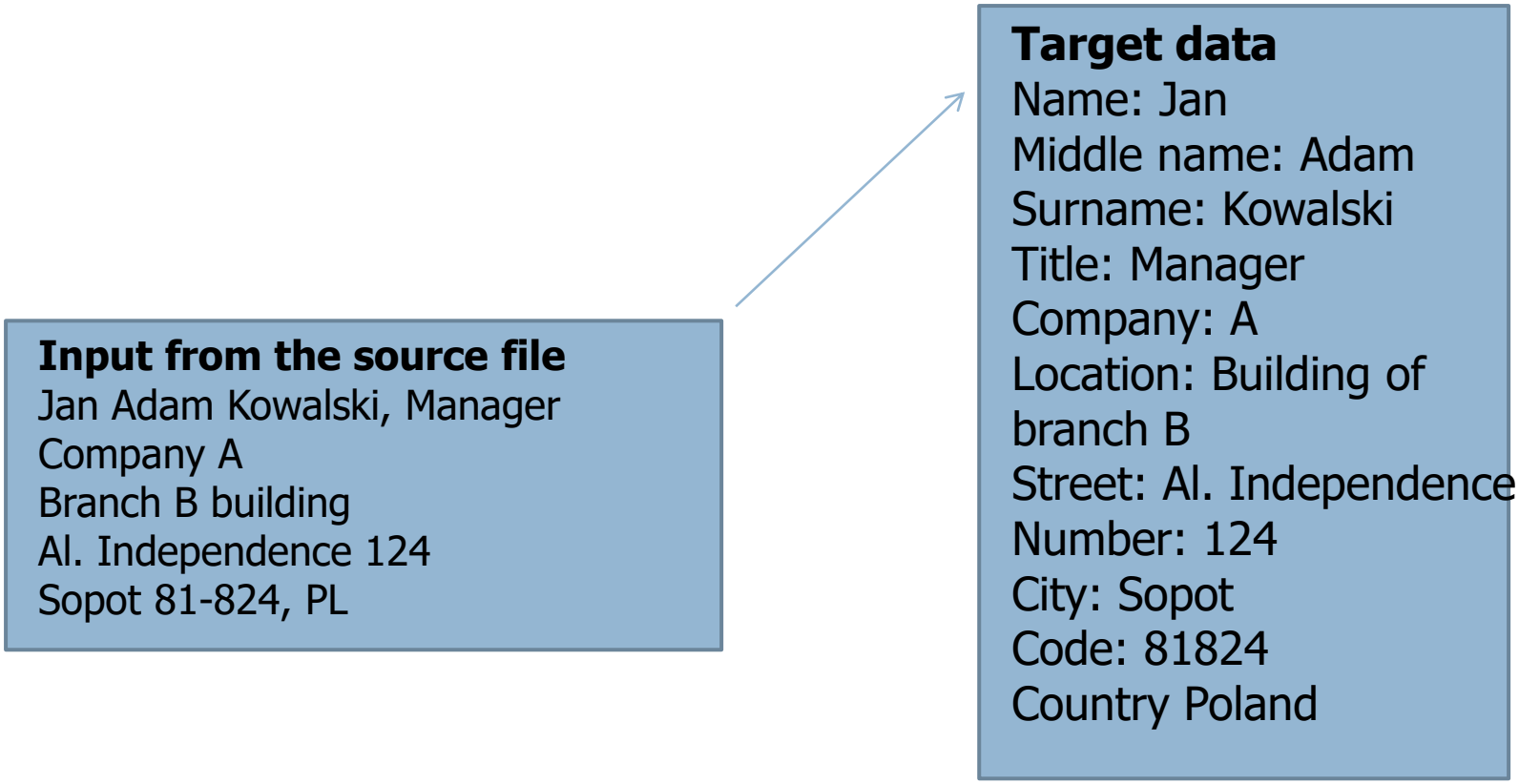
Data cleansing steps



Parsing

Parsing locates and identifies individual data elements in source files and then separates those data elements in target files.

Parsing


Input from the source file
Jan Adam Kowalski, Manager
Company A
Branch B building
Al. Independence 124
Sopot 81-824, PL
Target data
Name: Jan
Middle name: Adam
Surname: Kowalski
Title: Manager
Company: A
Location: Building of branch B
Street: Al. Independence
Number: 124
City: Sopot
Code: 81824
Country Poland

Correction

Correction is related to the parsing of individual data components through the use of complex algorithms and secondary data sources.

Correcting

Parsed data
Name: Jan
Middle name: Adam
Surname: Kowalski
Title: Manager
Company: A
Location: Building of branch B
Street: Al. Independence
Number: 124
City: Sopot
Code: 81824
Country Poland

Data corrected
Name: Jan
Middle name: Adam
Surname: Kowalski
Title: Manager
Company: A
Location: Building of branch B
Company address: …
Company PKD: …
Street: Al. Independence
Number: 124
City: Sopot
Code: 81824
Country Poland

Standardization

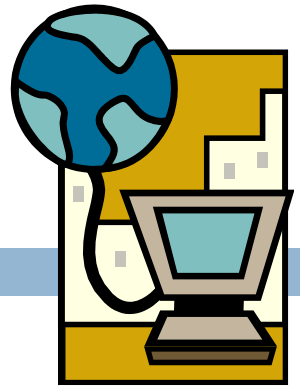
Standardization uses routine conversions to transform data to a preferred and consistent format, using standard and individual business rules.

Standardization

Data corrected
Name: Jan
Middle name: Adam
Surname: Kowalski
Title: Manager
Company: A
Location: Building of branch B
Company address: Gdańsk
NACE of the company: 37D
Street: Al. Independence
Number: 124
City: Sopot
Code: 81824
Country Poland

Standardized data
Prefix: Mr.
Name: Jan
Middle name: Adam
Surname: Kowalski
Title: Manager
Company: A
Location: Building of branch B
Company address: Gdańsk
NACE of the company: 37D
Street: Niepodległości
Number: 124
City: Sopot
Code: 81824
Country: PL

Matching

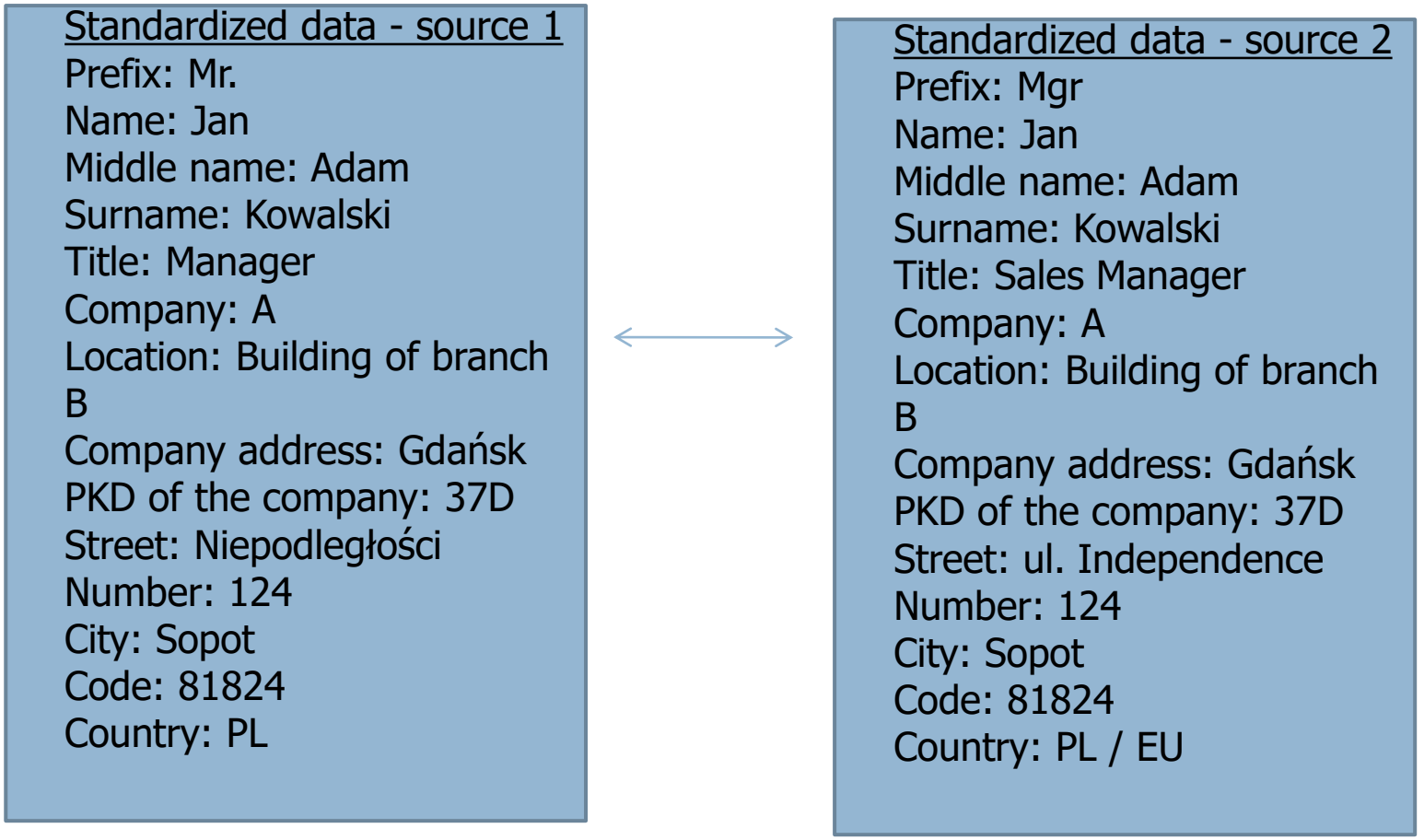


Find and match records within and between parsed, revised and standardized data, based on predefined business rules to eliminate duplicates.

Matching

Name	Street	Type of product	Client	City	Product code	Pattern	ID
						AAAAAA	P110
	-		-		NULL	ABAAA-	P115
	-		NULL			ABA-AA	P120
	-	-	-			ABCCAA	S300
-	-		-			BBACAA	S310

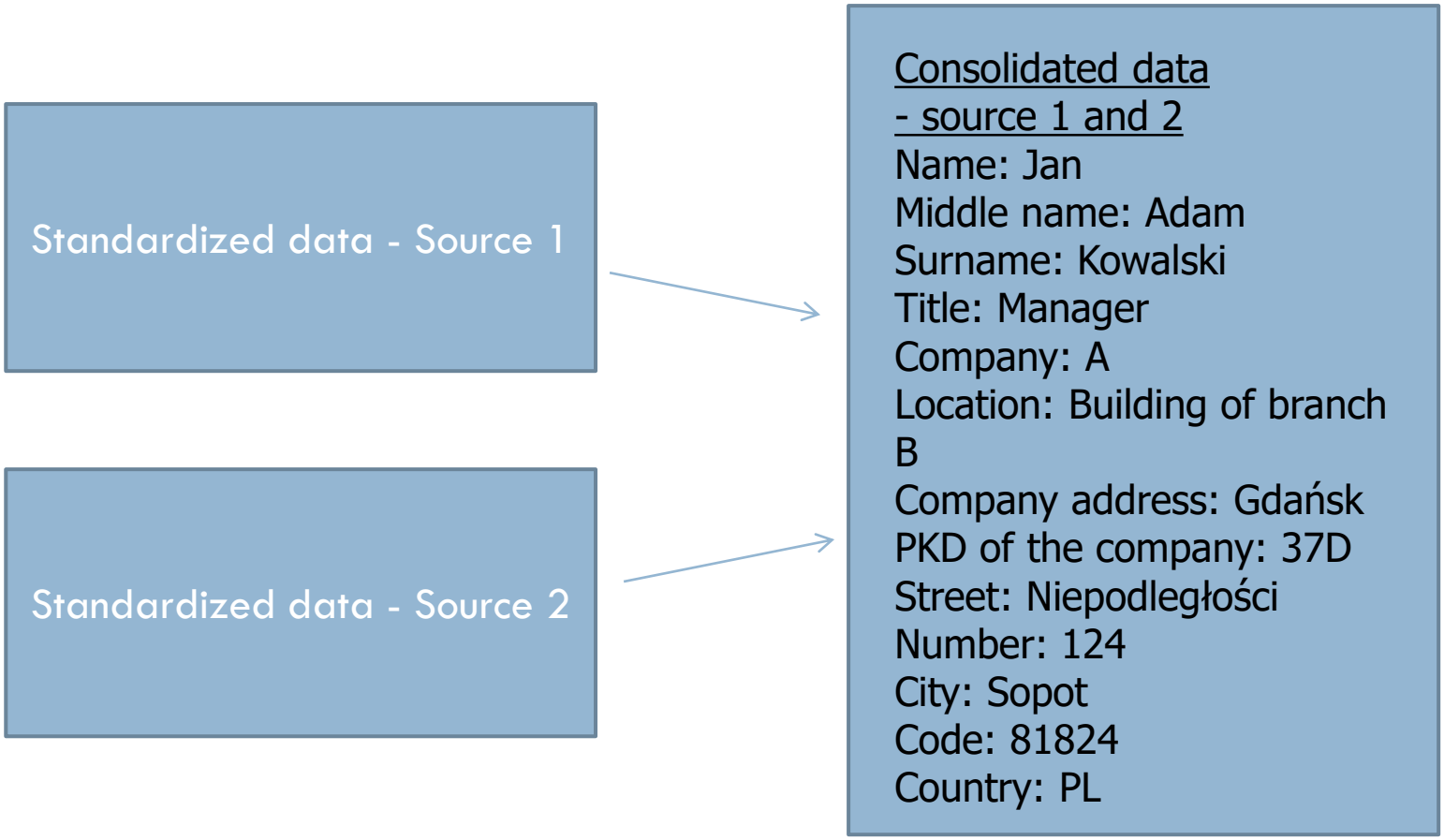
Matching


Standardized data - source 1
Prefix: Mr.
Name: Jan
Middle name: Adam
Surname: Kowalski
Title: Manager
Company: A
Location: Building of branch B
Company address: Gdańsk
PKD of the company: 37D
Street: Niepodległości
Number: 124
City: Sopot
Code: 81824
Country: PL
Standardized data - source 2
Prefix: Mgr
Name: Jan
Middle name: Adam
Surname: Kowalski
Title: Sales Manager
Company: A
Location: Building of branch B
Company address: Gdańsk
PKD of the company: 37D
Street: ul. Independence
Number: 124
City: Sopot
Code: 81824
Country: PL / EU

Consolidation

Analyze and identify relationships between records and merge them into a single whole.

Consolidation



Question

What method of integration is concerned with storing data exclusively in source databases and accessing them using perspectives?

- schema integration
- virtual data integration
- materialized data integration
- integration of thematic wholesalers