IRIT
Institut de Recherche
en Informatique de Toulouse
CNRS - INP - UT3 - UT1 - UT2J

eric
ENTREPÔTS, REPRÉSENTATION
& INGÉNIERIE des CONNAISSANCES

WP 2 : Automatic data integration in multidimensional data warehouses

Yuzhao YANG
Franck RAVAT
Olivier TESTE
Jérôme DARMONT

Contents

01 **Introduction**

02 **Automatic Data Warehouse Design**

03 **Data Warehouse Merging**

04 **Data Warehouse Imputation**

05 **Implementation**

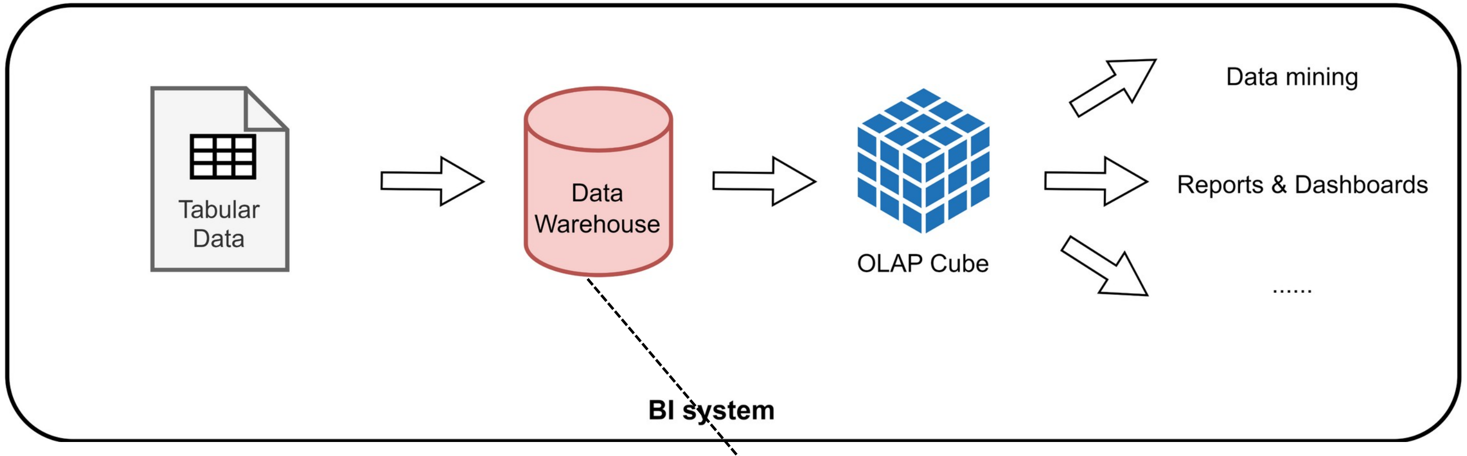
06 **Conclusion**

01 Introduction

Research Context

Large Companies ✓

Small Entities ✗ Lack of budget and experts
✓


Tabular Data

Data Warehouse

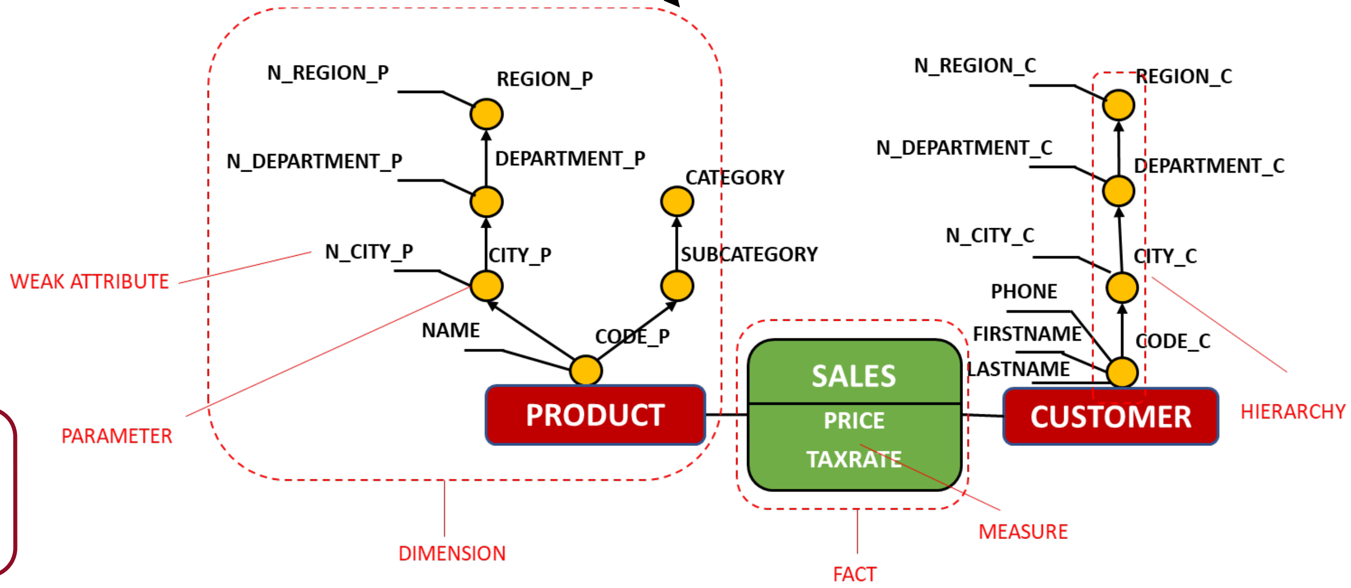
OLAP Cube

Data mining

Reports & Dashboards

......

BI system

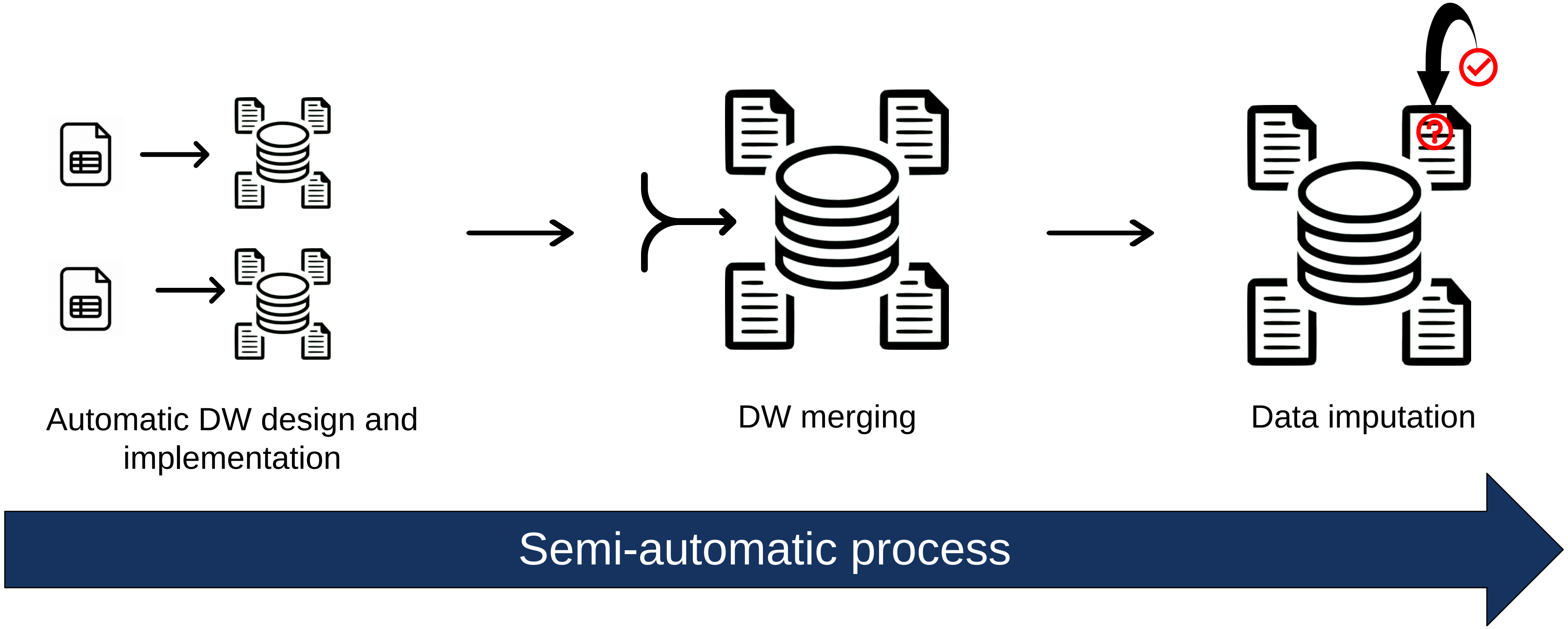


BI4people (Business intelligence for the people)

How can we automatically integrate tabular data integration in multidimensional data warehouses?

01 Introduction

Solution



Automatic DW design and implementation

DW merging

Data imputation

Semi-automatic process

Contents

01 Introduction

02 **Automatic Data Warehouse Design**

03 Data Warehouse Merging

04 Data Imputation

05 Implementation

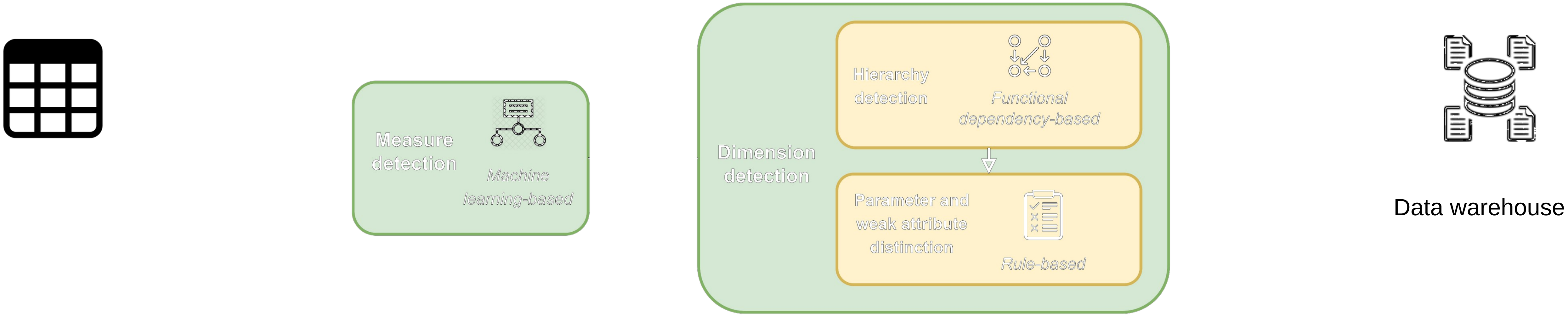
06 Conclusion

02 Automatic DW Design

Problems

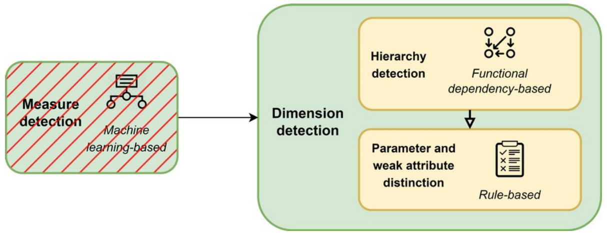
How to automatically generate a DW from tabular data?

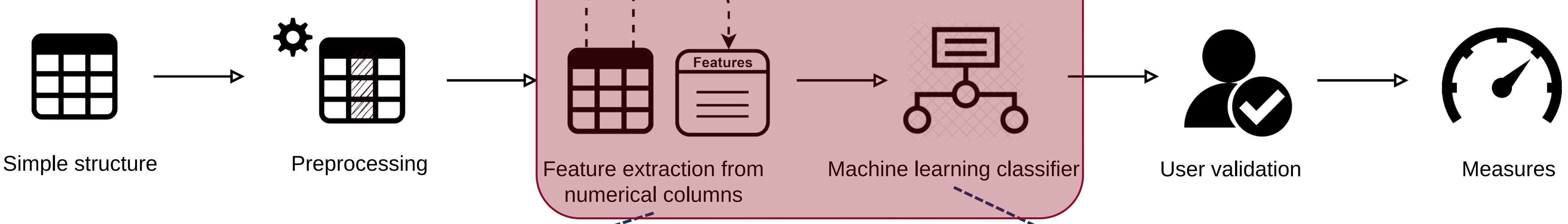
- **Lack of schema**
- **Complex DW structure**



02 Automatic DW Design

Measure Detection




Features
Simple structure
Preprocessing
Feature extraction from numerical columns
Machine learning classifier
User validation
Measures

Feature Category	Feature
General feature	Data Type
	Positive/Negative/Zero value ratio
	Unique value ratio
Statistical feature	Same digital number
	Average/Minimum/Maximum/Median/Upper quartile/Lower quartile values
	Coefficient of variation
	Range ratio
Inter-column feature	Location ratio
	Numerical column ratio
	Multiple functional dependencies
	Numerical neighbor

- **Support vector machine (SVM)**
- **Decision tree (DT)**
- **Random forest (RF)**
- **K-nearest neighbors (KNN)**

Random forest

- Best F-score
- Stable distribution

Contents

01 Introduction

02 Automatic Data Warehouse Design

03 **Data Warehouse Merging**

04 Data Imputation

05 Implementation

06 Conclusion

03 DW Merging

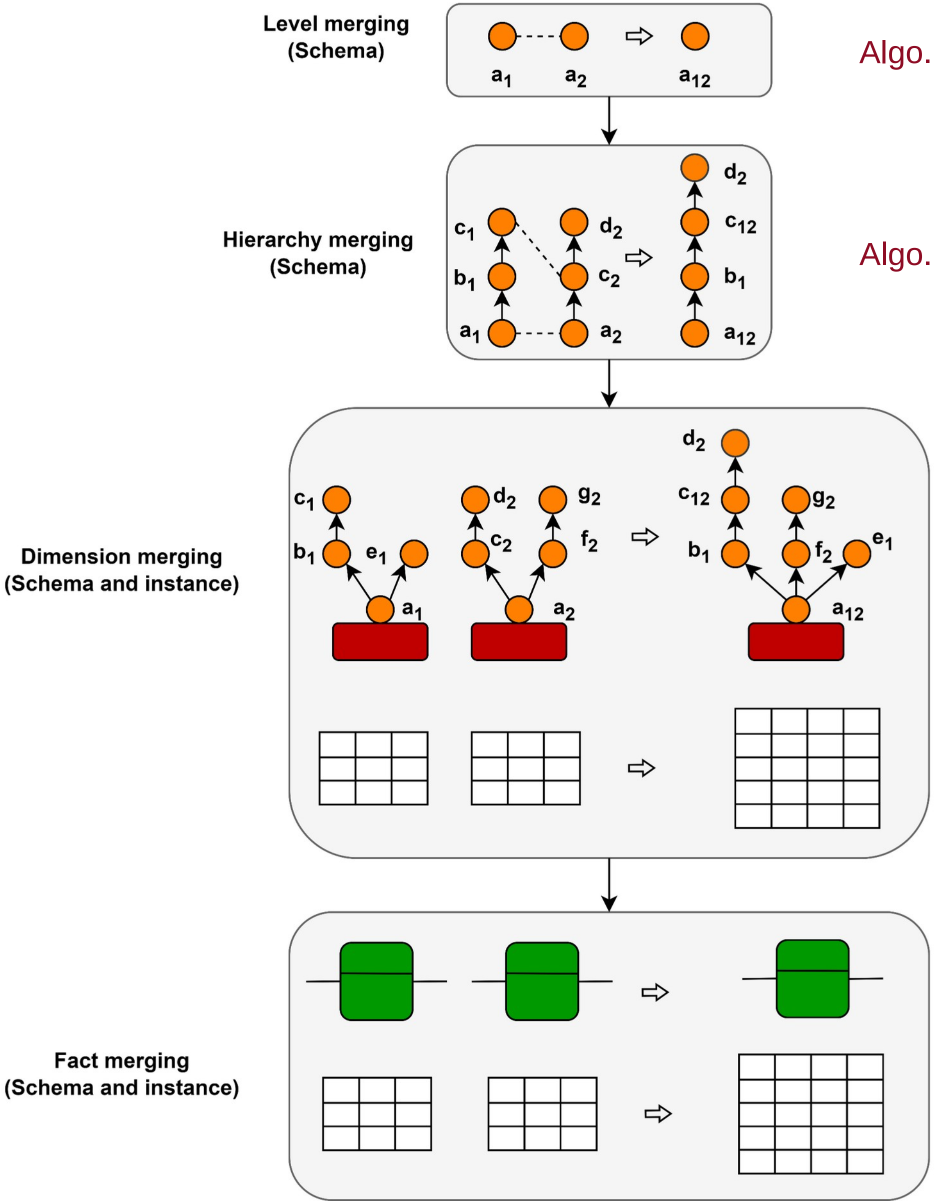
Problems

How to merge two DWs having commun elements?

- Merging at both schema and instance levels
- Generation of different types of schema (star, constellation)

03 DW Merging

Contribution



Level merging (Schema)

Algo. 5 mergeHierarchies, p68

Hierarchy merging (Schema)

Algo. 5 mergeHierarchies, p68

Dimension merging (Schema and instance)

Algo. 6 mergeDimensions, p73

Fact merging (Schema and instance)

Algo. 8 mergeStars, p78

Content

01 Introduction

02 Automatic Data Warehouse Design

03 Data Warehouse Merging

04 **Data Imputation**

05 Implementation

06 Conclusion

04 Data Imputation

Problems

How to carry out data imputation to ensure consistent analysis?

- **Dimension**
- **Categorical**

- **Deducted values**
- **Predicted values**

Hie - OLAPKNN

Hierarchical imputation → **OLAPKNN imputation**

Functional dependencies in hierarchies

Non-parametric and instance-based

Suitable for different types of data

Relatively high accuracy

✓ Deducted values

✗ Limited

✓ Specific distance metric

✓ Consideration of dependencies

06 Conclusion

Contribution

Contributions on Automatic DW Design from Tabular Data

- **Mesure Detection**
  - *Machine learning classification*
  - *Random forest : +17%*
  - *Relevant features*
  - *Generic model*

- **Dimension Detection**
  - *Hierarchy: functional dependency*
  - *Parameter and weak attribute: rules*
  - *Dimension: 100%*
  - *Hierarchy: 67% - 100%*

Contributions on Automatic DW Merging

- **DW merging**
  - *Schema and instance*
  - *Generation of star or constellation schema*

06 Conclusion

Contribution

Contributions on Data Imputation

- **Hie-OLAPKNN**
    - *Hierarchical imputation*
    - *OLAPKNN: specific distance*
    - *Effective : + 45%*
    - *Efficient :* -+19 times

Contributions on Tabular Data Integration Application

- **Application**
    - *3 fonctionnalities*
    - *User-friendly interface*
    - *Non-expert and expert version*

Publications

- **Automatic DW Design from Tabular Data**
  - Yuzhao Yang, Fatma Abdelhédi, Jérôme Darmont, Franck Ravat, Olivier Teste: Automatic Machine Learning-Based OLAP Measure Detection for Tabular Data. **DaWaK** 2022: 173-188
  - Yuzhao Yang, Jérôme Darmont, Franck Ravat, Olivier Teste: Automatic Integration Issues of Tabular Data for On-Line Analysis Processing. **EDA** 2020: 5-18
- ***Automatic DW Merging***
  - Yuzhao Yang, Jérôme Darmont, Franck Ravat, Olivier Teste: An Automatic Schema-Instance Approach for Merging Multidimensional Data Warehouses**.** **IDEAS** 2021: 232-241
- ***Data Imputation***
  - *Yuzhao Yang, Jérôme Darmont, Franck Ravat, Olivier Teste: Dimensional Data KNN-Based Imputation.* ***ADBIS*** *2022: 315-329*
  - *Yuzhao Yang, Fatma Abdelhédi, Jérôme Darmont, Franck Ravat, Olivier Teste: Internal Data Imputation in Data Warehouse Dimensions.* ***DEXA*** *(1) 2021: 237-244*

PhD Thesis

- **Tabular data integration for multidimensional data warehouse**
  (https://theses.fr/2022TOU10052)

06 Conclusion

Future work

Short-term plan

- Imputation by External Sources

Mid-term plan

- Schema Evolution of Sources

Long-term plan

- Imputation Algorithm Generalization for ML algorithms
- Data Science with Data Lake
  - Merging / Matching Datasets
  - Imputing Data
  - User-friendly DL

Thank you!