UNIVERSITY
OF LATVIA

Data warehouse creation methods in big data era

Jānis Zemnickis

Supervisor: Dr. sc. comp., prof. Laila Niedrīte

2024

Actuality

- Data warehouses are widely used in organizations, and they are one of the most powerful analytical solutions

- Data warehouse creation methods, data models, and architecture has changed in the era of big data and NoSQL databases

- Topic is presented in International Conference on Enterprise Information Systems, World Conference on Information Systems and Technologies, International Baltic Conference on Digital Business and Intelligent Systems, etc.

Problems

1. Data warehouses fail because they are not developed according to stakeholders’ needs

2. Traditional data warehouse data models, architecture, creation methods can not provide all needs in context of growing data amounts

3. There is unused unstructured data potential and lack of usage in organizations

Goal

Develop new methods to improve the data warehouse requirement engineering phase and extend the data warehouse data model by using unstructured data and NLP technologies.

Tasks

1. To investigate the data warehouse architecture, data warehouse data model development methods and their evolution in the context of big data;
2. To conduct literature research on traditional data warehouse requirements acquisition methods;
3. Explore the possibilities of formalising the information requirements of a data warehouse;
4. Conduct literature research and analysis on requirements engineering activities in big data development projects;
5. Explore natural language processing techniques and tools, evaluate the potential of natural language as a data source;
6. Examine the concept and working principles of organisation's KPIs;
7. Propose a new or improve an existing method for generating data warehouse pre-schemas using formal data warehouse requirements;
8. Propose new methods to enable new KPIs generation and to extend the data warehouse model by analysing unstructured data;
9. To perform approbation of methods by improving the data warehouse data model and defining KPIs for the organisation.

Terms

- Data warehouse - is a subject oriented, integrated, time-variant, and non-volatile collection of data
- Big data – described by its characteristics - volume, velocity, variety
- Big data era – time when organizations start rapidly invest in data (2005)
- NLP - Natural Language Processing - reveal the structure and meaning of text
- KPI - Key Performance Indicator. KPI defines a set of actions which achieving organization performance will be increased significantly

Data warehouse creation methods in big data era

1. Data driven – based on source systems schemas and available data
   - Data warehouses based on NoSQL databases

     a) Document-oriented (*CHEVALIER, Max, et al. Implementing multidimensional data warehouses into NoSQL*)

     b) Column based (*FERREIRA, Leandro Mendes; ALVES-SOUZA, Solange Nice; DA SILVA, Luciana Maria. Startable: Multidimensional Modelling for Column-Oriented NoSQL*)
   - Data warehouses based on graph databases (*SELLAMI, Amal; NABLI, Ahlem; GARGOURI, Faiez. Graph NoSQL data warehouse creation*)

2. Requirement driven - based on stakeholder requirements
   - Data warehouses creation based on pivot tables (*BIMONTE, Sandro; ANTONELLI, Leandro; RIZZI, Stefano. Requirements-driven data warehouse design based on enhanced pivot tables*)

UNIVERSITY OF LATVIA

Data warehouse requirement gathering in big data era * **

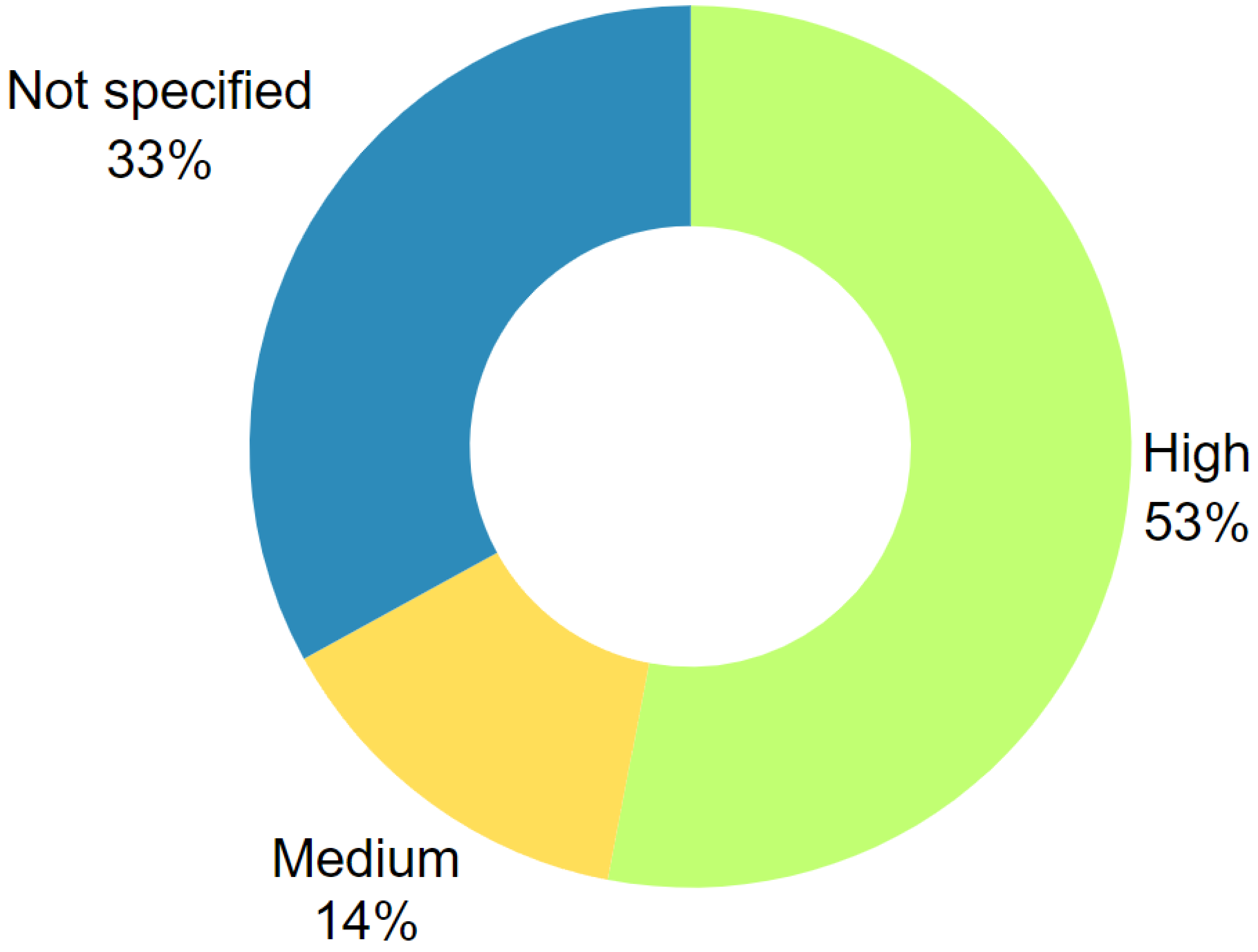
- Research to explore existing big data processing methods in context of data warehouse requirement gathering

- Analyzed 22 big data processing methods by common aspects:
    - Which requirement engineering activities are in focus
    - Which requirement engineering artifacts are applied
    - Applicability of requirement development activities in Big data context
    - Etc.

* KOZMINA, Natalija; NIEDRITE, Laila; **ZEMNICKIS, Janis**. Information requirements for big data projects: A review of state-of-the-art approaches. In: Databases and Information Systems: 13th International Baltic Conference, DB&IS 2018, Trakai, Lithuania, July 1-4, 2018, Proceedings 13. Springer International Publishing, 2018. p. 73-89.

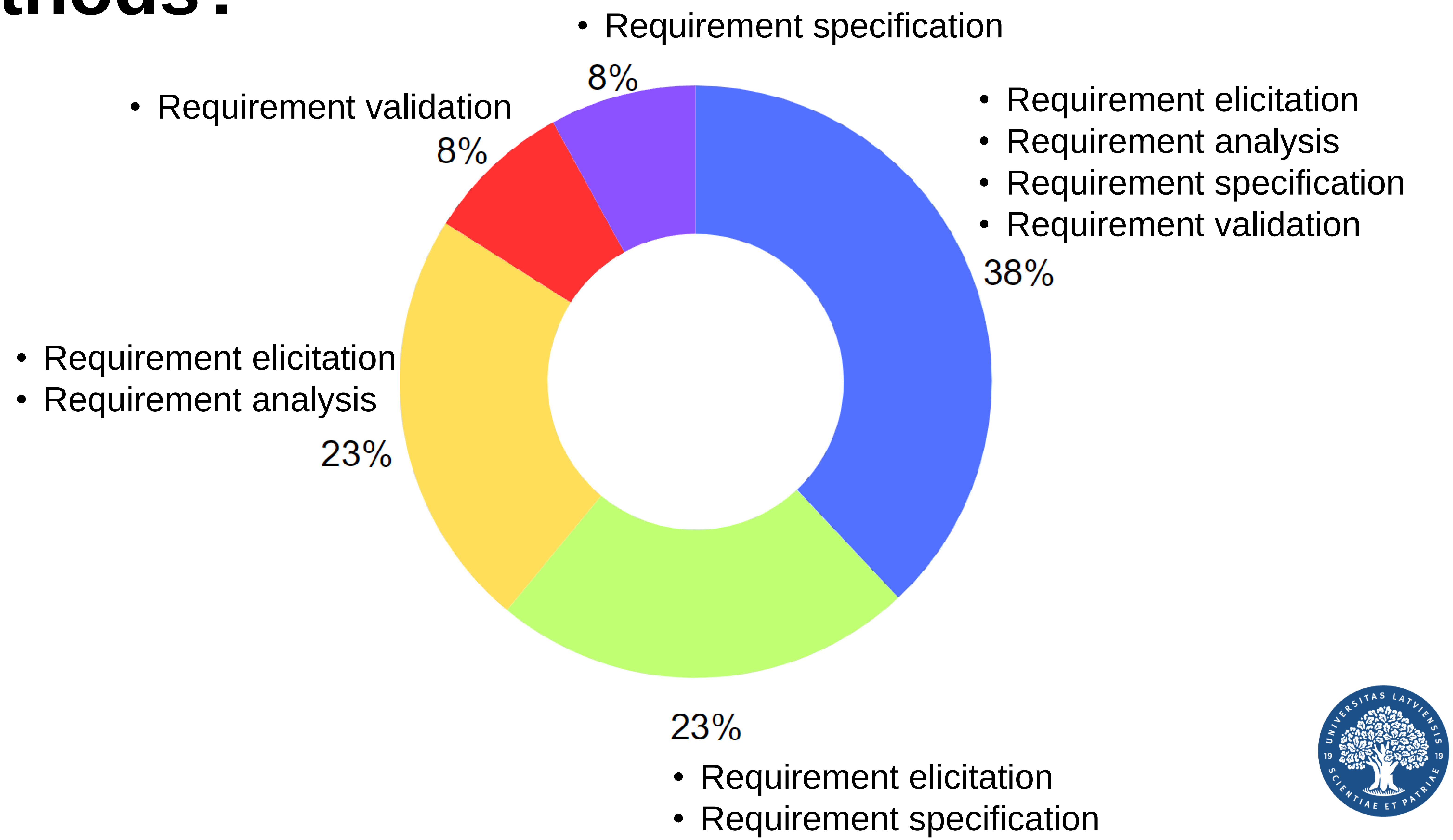
** KOZMINA, Natalija; NIEDRITE, Laila; **ZEMNICKIS, Janis**. Perspectives of Information Requirements Analysis in Big Data Projects. In: Databases and Information Systems X. IOS Press, 2019. p. 109-124.

Applicability of DW requirement engineering activities using big data processing methods



- Requirements engineering activities:
  - Requirement elicitation
  - Requirement analysis
  - Requirement specification
  - Requirement validation

Which data warehouse requirement engineering activities can be applied using big data processing methods?



UNIVERSITY OF LATVIA

Big data processing methods - purposes

- To analyze information system user behaver *(FERNANDEZ-GARCIA, Antonio Jesus, et al. Evolving mashup interfaces using a distributed machine learning and model transformation methodology)*

- To analyze text with improved approach *(CHEPTSOV, Alexey, et al. Introducing a new scalable data-as-a-service cloud platform for enriching traditional text mining techniques by integrating ontology modelling and natural language processing)*

- To improve data management and visualization solution *(TARDIO, Roberto; MATE, Alejandro; TRUJILLO, Juan. An iterative methodology for big data management, analysis and visualization)*

- It is possible to use Big data processing methods for data warehouse requirement engineering but none of analyzed methods currently defines to do that

UNIVERSITY OF LATVIA

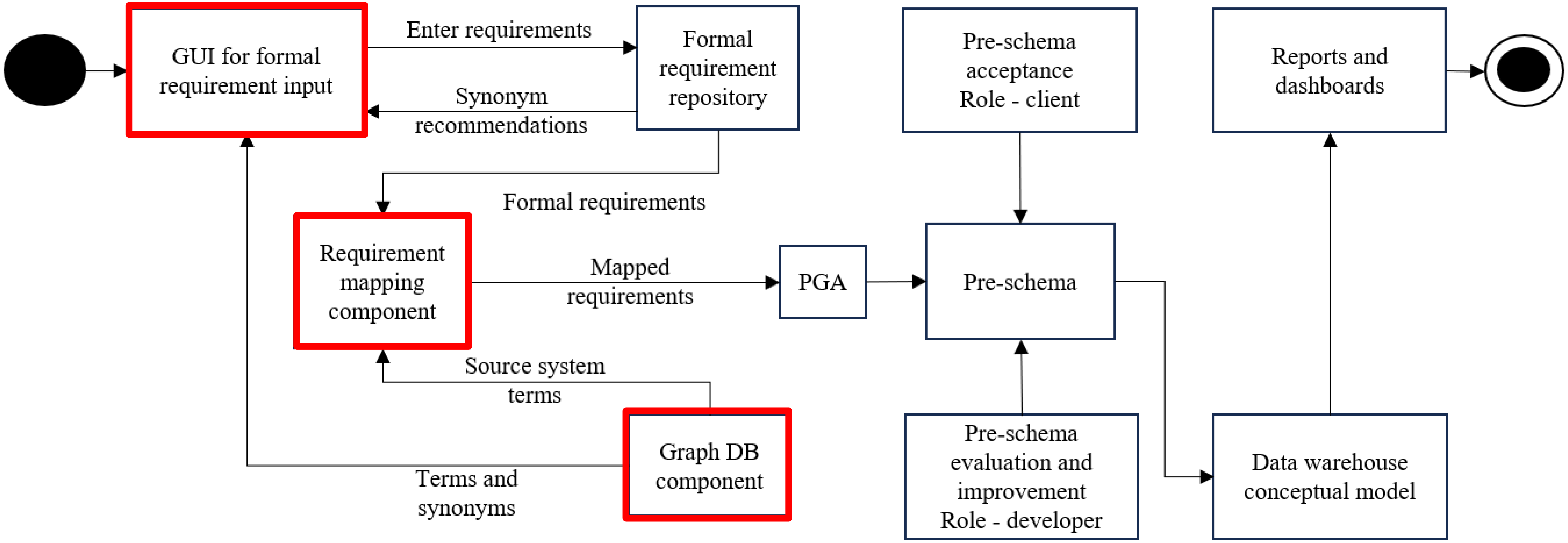
Free text as an information source

- Unstructured data analysis is more complex than structured data analysis (*KASSNER, Laura, et al. Product life cycle analytics–next generation data analytics on structured and unstructured data. Procedia CIRP, 2015, 33: 35-40*)

- Information from unstructured data can be useful and interesting for the decision-making process (*Pejić Bach, M., Krstić, Ž., Seljan, S., Turulja, L. (2019). Text mining for big data analysis in financial sector: A literature review. Sustainability*)

- 95% of organisations struggle with effective unstructured data analysis (*Baviskar, D., Ahirrao, S., Kotecha, K. (2021). Multi-layout unstructured invoice documents dataset: A dataset for template-free invoice processing and its evaluation using AI approaches*)

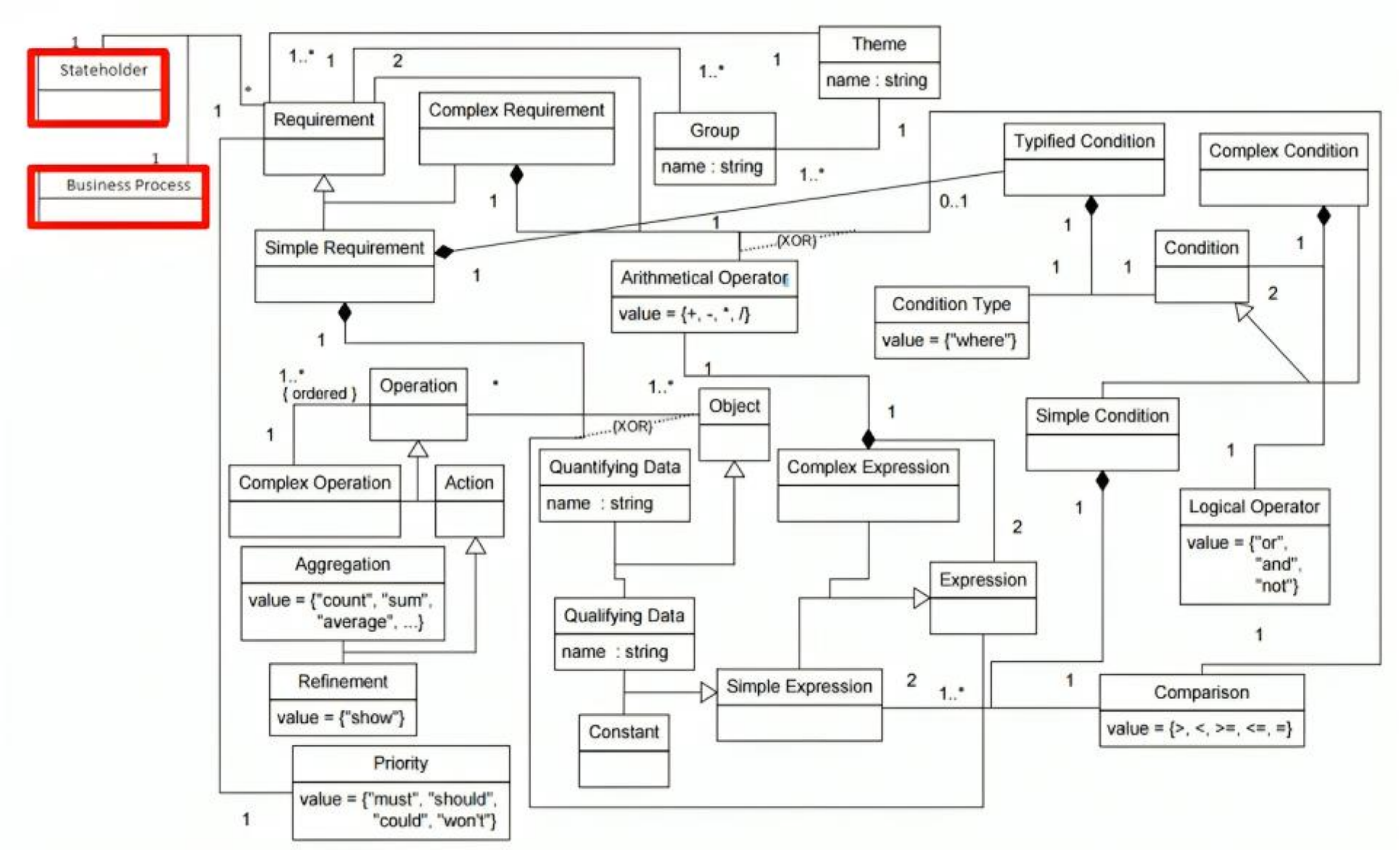
Method to generate conceptual data warehouse data model *


GUI for formal requirement input
Enter requirements
Formal requirement repository
Synonym recommendations
Pre-schema acceptance Role - client
Reports and dashboards
Formal requirements
Requirement mapping component
Mapped requirements
PGA
Pre-schema
Source system terms
Graph DB component
Terms and synonyms
Pre-schema evaluation and improvement Role - developer
Data warehouse conceptual model

* KOZMINA, Natalija; NIEDRITE, Laila; **ZEMNICKIS, Janis**. Gathering Formalized Information Requirements of a Data Warehouse. In: International Conference on Enterprise Information Systems. SCITEPRESS, 2017. p. 217-224.

UNIVERSITY OF LATVIA

Improved formal data warehouse requirements metamodel

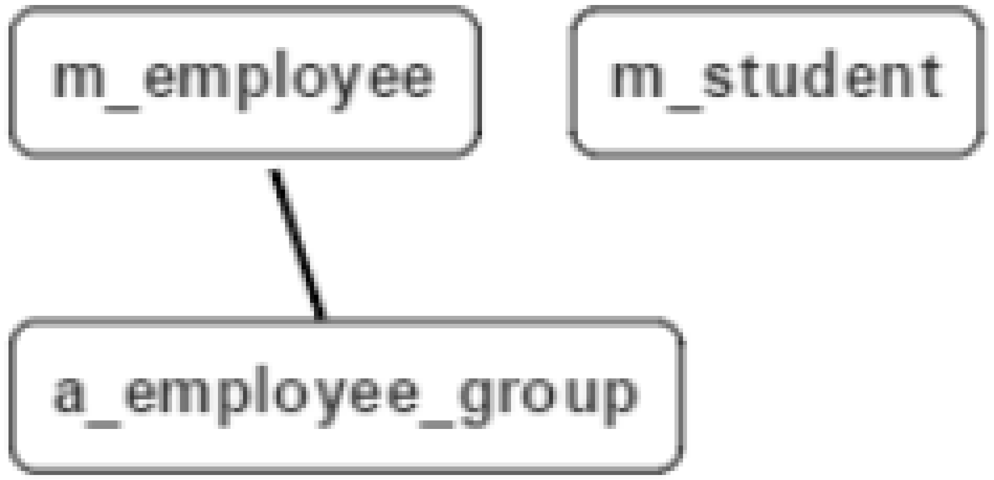


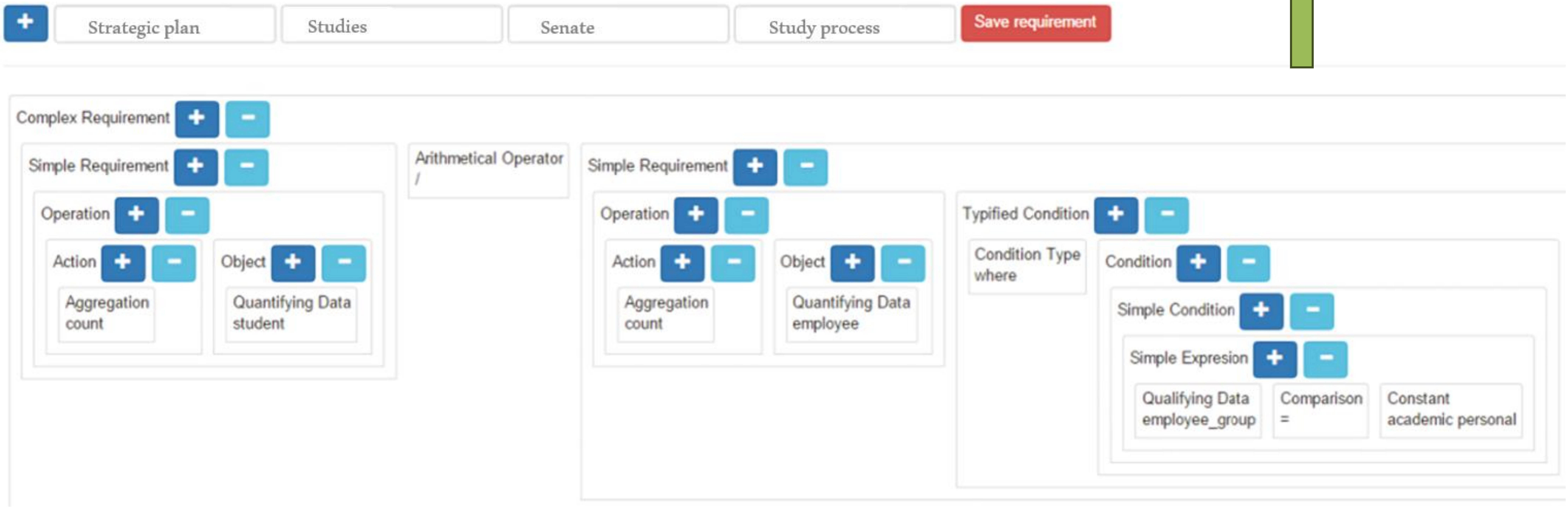
UNIVERSITY OF LATVIA

Example of the method in action

Information requirement: “Show student and academic personal ratio”.

Multidimensional data model elements





Complex Requirement + −

Simple Requirement + −

Operation + −

Action + −

Aggregation
count

Object + −

Quantifying Data
student

Arithmetical Operator
/

Simple Requirement + −

Operation + −

Action + −

Aggregation
count

Object + −

Quantifying Data
employee

Typified Condition + −

Condition Type
where

Condition + −

Simple Condition + −

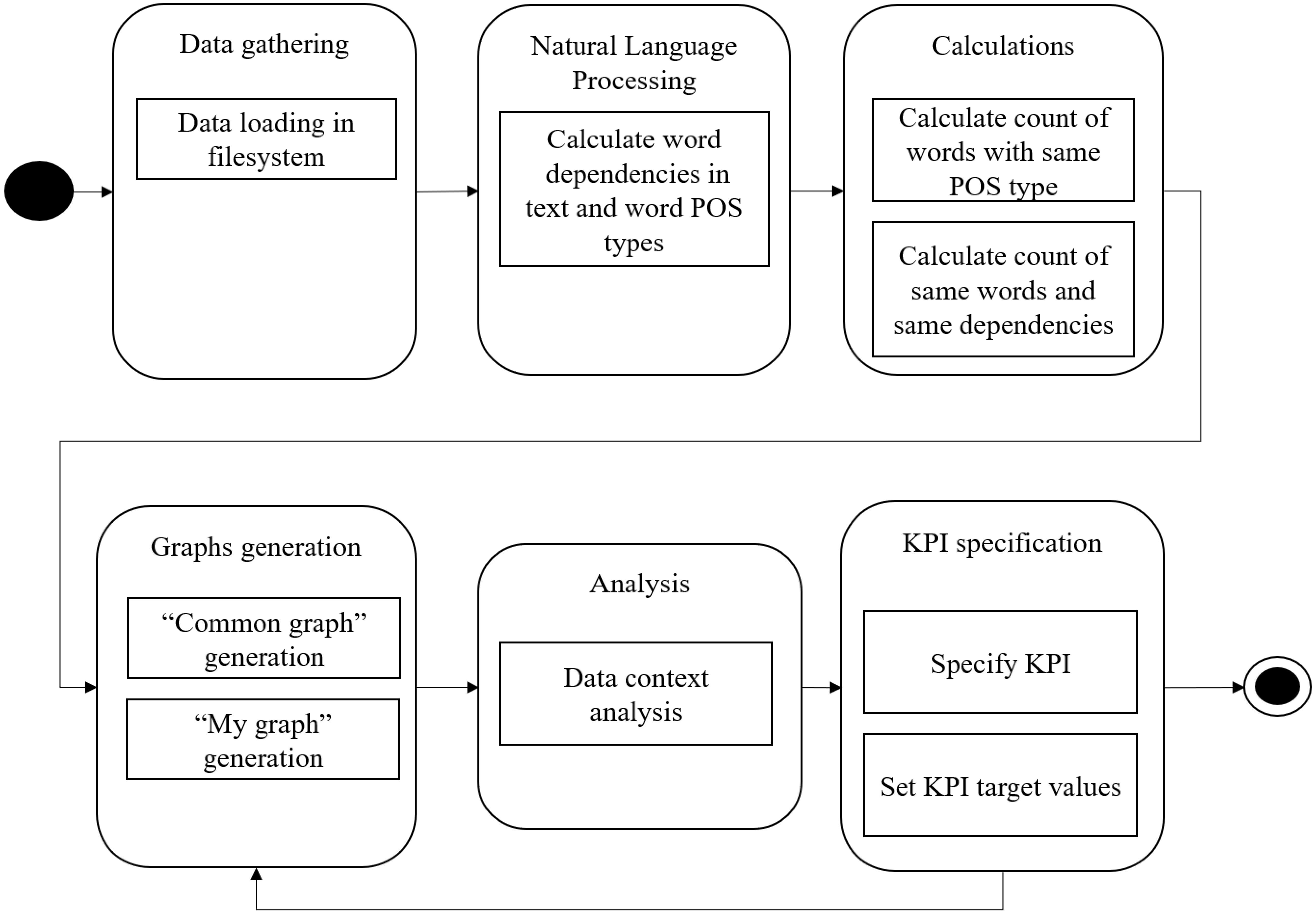
Simple Expresion + −

Qualifying Data
employee_group

Comparison
=

Constant
academic personal

Method to define KPIs using unstructured text *



Data gathering

Data loading in filesystem

Natural Language Processing

Calculate word dependencies in text and word POS types

Calculations

Calculate count of words with same POS type

Calculate count of same words and same dependencies

Graphs generation

"Common graph" generation

"My graph" generation

Analysis

Data context analysis

KPI specification

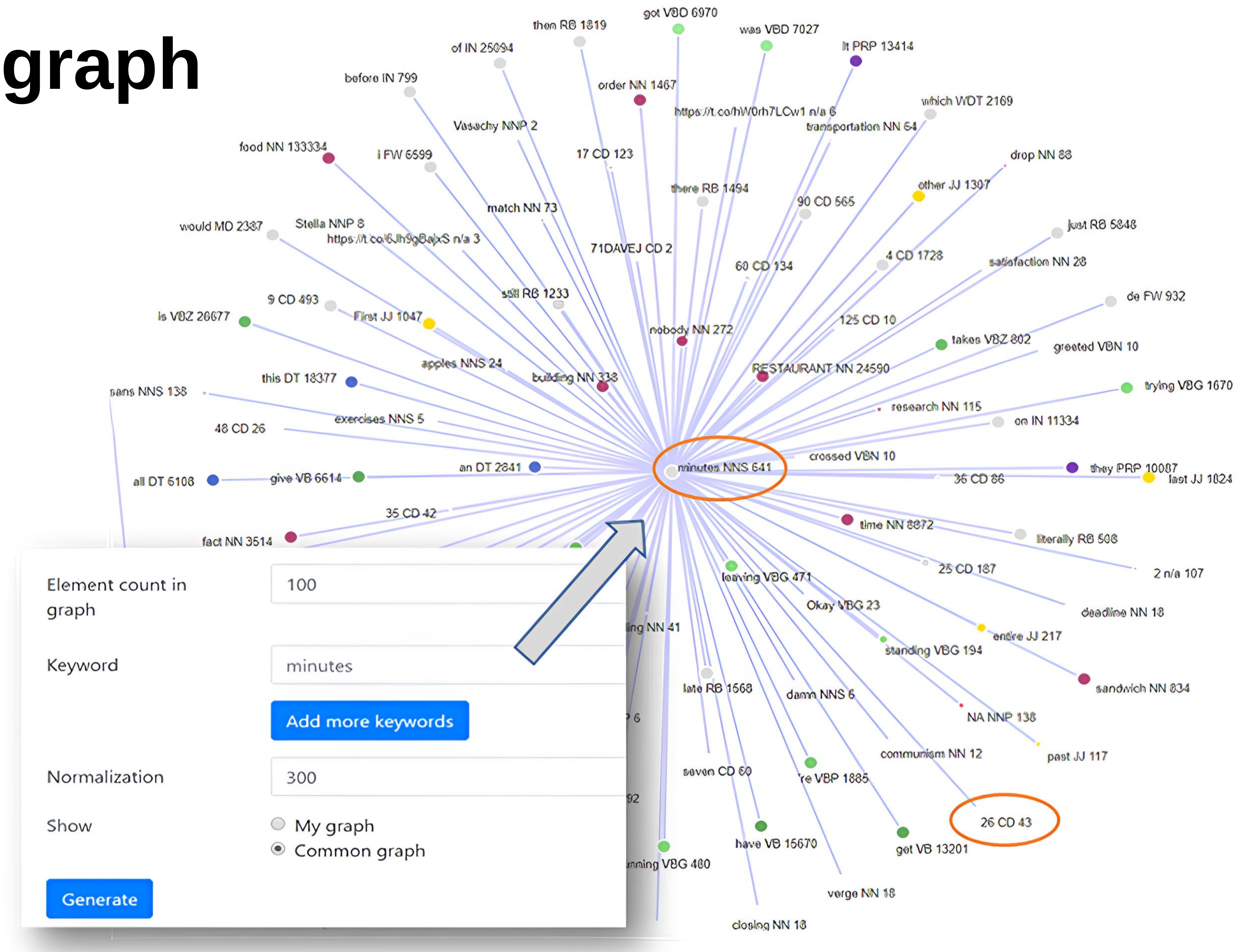
Specify KPI

Set KPI target values

* **ZEMNICKIS, Janis**; NIEDRITE, Laila; KOZMINA, Natalija. A Little Bird Told Me: Discovering KPIs from Twitter Data. In: Databases and Information Systems: 14th International Baltic Conference, DB&IS 2020, Tallinn, Estonia, June 16–19, 2020, Proceedings 14. Springer International Publishing, 2020. p. 161-175.

UNIVERSITY OF LATVIA

Common graph



Element count in graph: 100

Keyword: minutes

Add more keywords

Normalization: 300

Show: ○ My graph ● Common graph

Generate

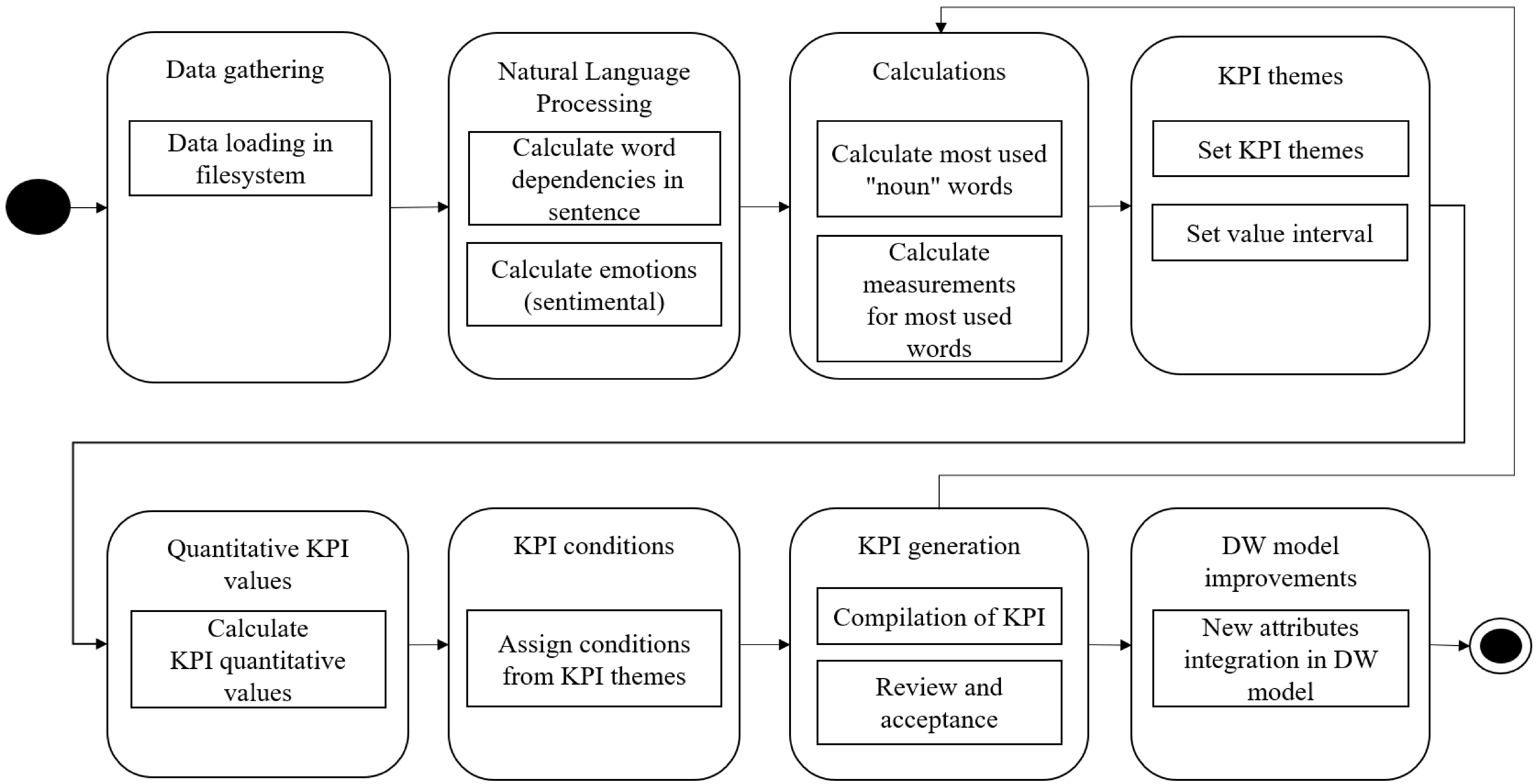
KPI specification

- Specify KPI according to taxonomy (Domínguez, E., Pérez, B., Rubio, Á. L., & Zapata, M. A. (2019). A taxonomy for key performance indicators management. Computer Standards & Interfaces, 64, 24-40.)
  1. Performance measurement (Financial, ***Customer***, Internal Process, and Learning and Growth) and Scope perspective (***Catering***)
  2. KPI hardness perspective (***Soft/Hard***)

Key word	Tweet
30	Wings to go took 30 minutes for 2 orders of wings and a kids quesadilla just took forget our sides. It takes a lot for me to get irritated with a food establishment but I'm annoyed to say the least
35	@PopeyesChicken been waiting 35 minutes for my food. So much for fast food. Disappointing.
5	Gluten-Free Vegan No-Cook Raw Cranberry Sauce...made with only 6 clean, real food ingredients and is ready in 5 minutes!

KPI: Wait time should be less then 30 minutes

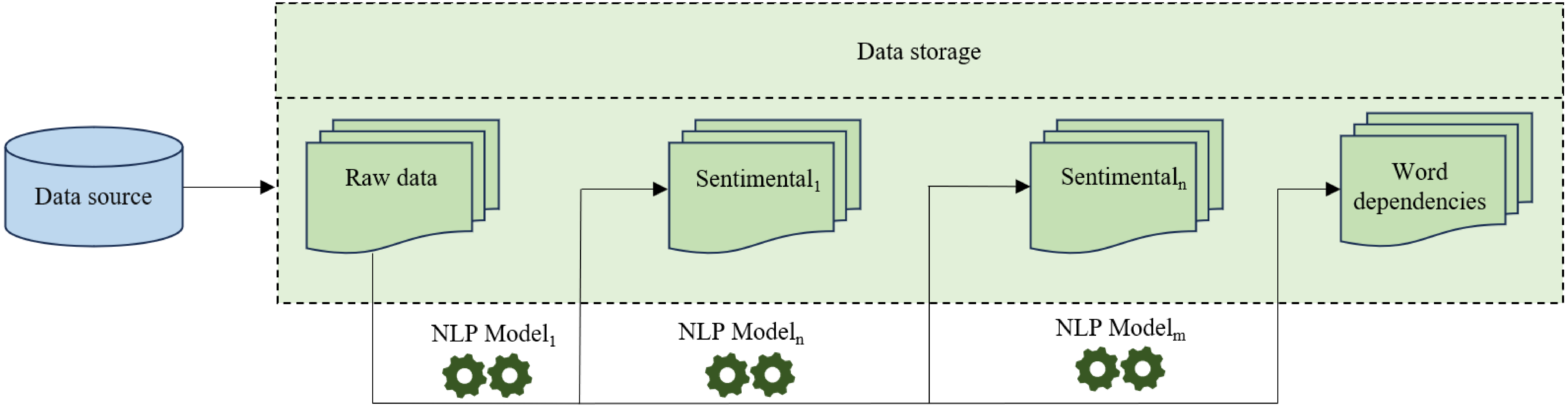
Method data warehouse data model improvements from customer feedback *


Data gathering
Data loading in filesystem
Natural Language Processing
Calculate word dependencies in sentence
Calculate emotions (sentimental)
Calculations
Calculate most used "noun" words
Calculate measurements for most used words
KPI themes
Set KPI themes
Set value interval
Quantitative KPI values
Calculate KPI quantitative values
KPI conditions
Assign conditions from KPI themes
KPI generation
Compilation of KPI
Review and acceptance
DW model improvements
New attributes integration in DW model

* **ZEMNICKIS, Jānis**. Data Warehouse Data Model Improvements from Customer Feedback. *Baltic Journal of Modern Computing*, 2023, 11.3.

UNIVERSITY OF LATVIA

Data loading and natural language processing



Twitter API

hadoop HDFS

John Snow LABS

APACHE Spark

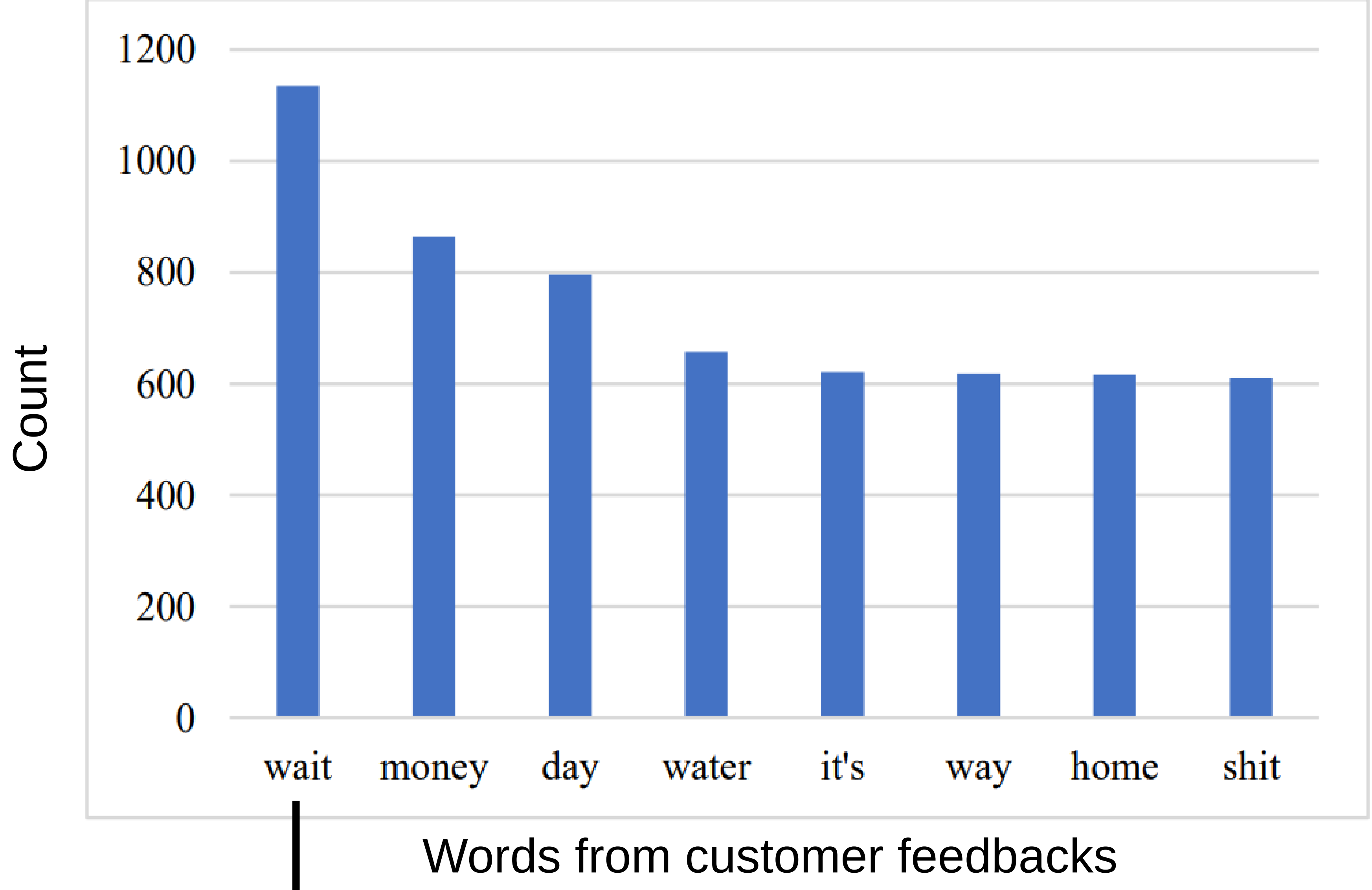
PySpark

jupyter

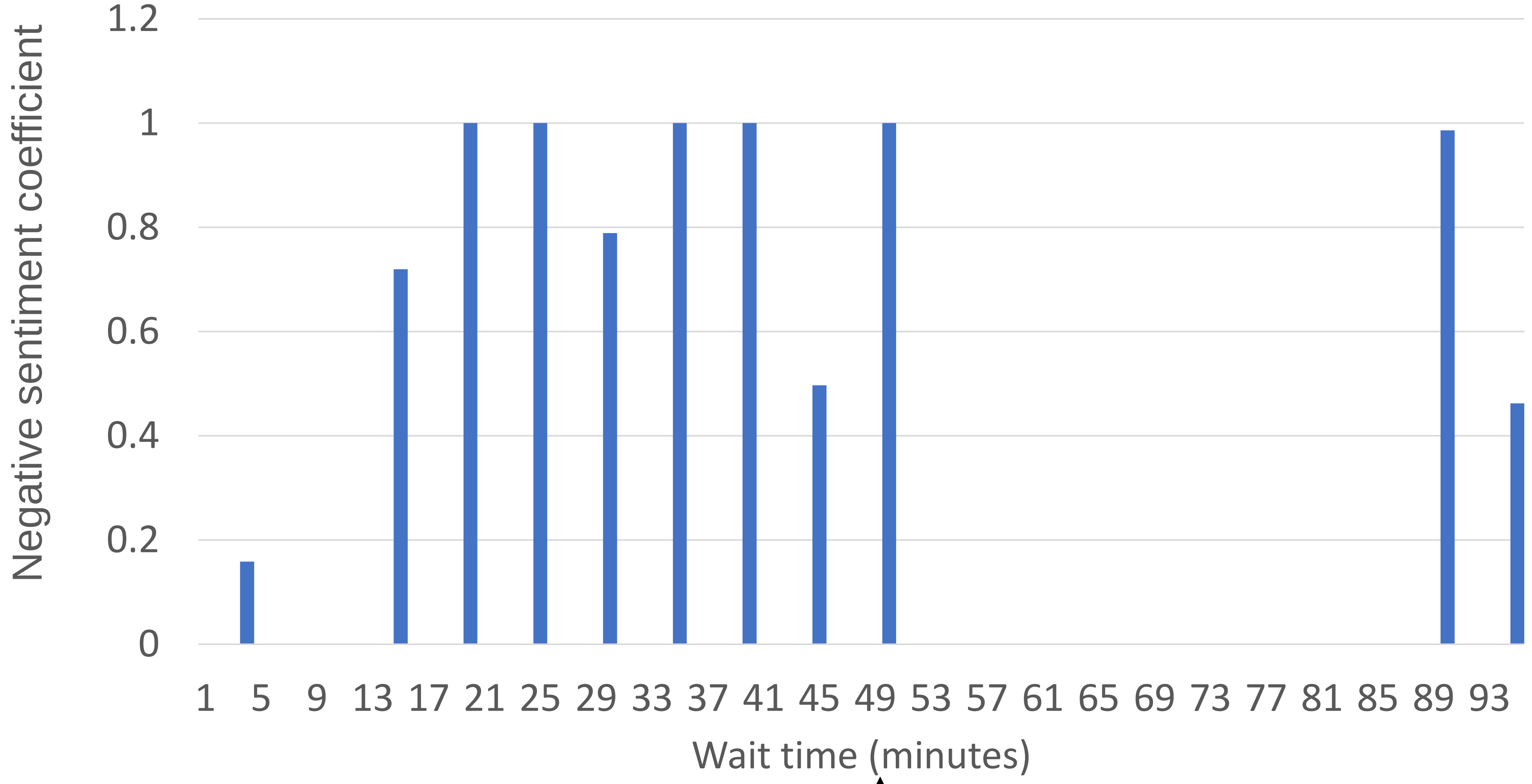
UNIVERSITY OF LATVIA

Data calculations

Most used words

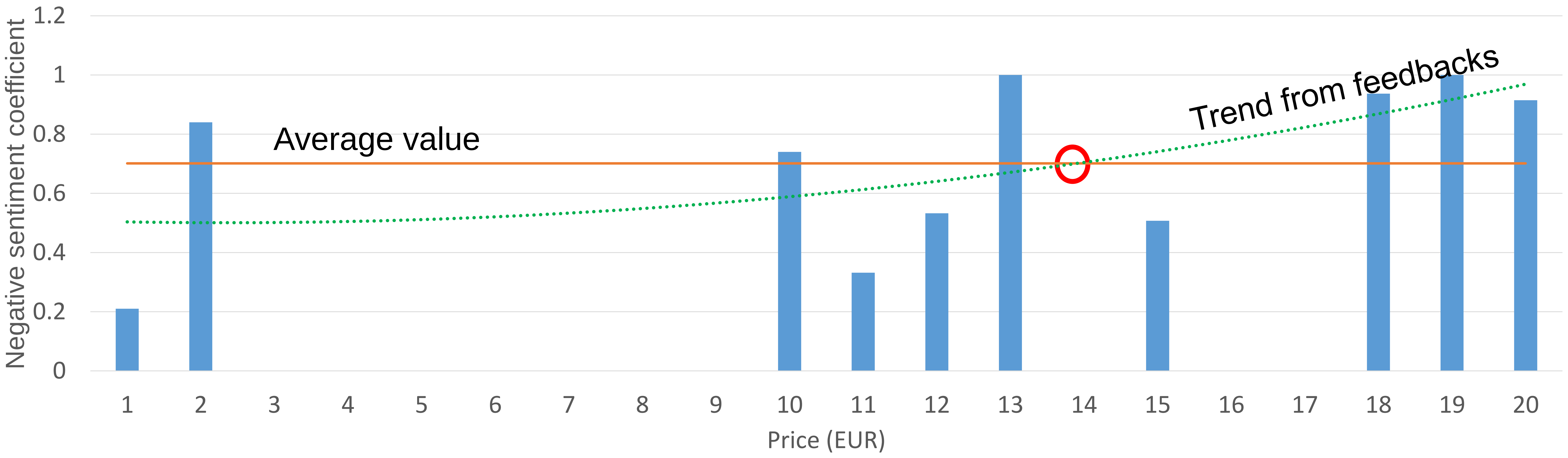


Most frequent nouns related to measurement



KPI target value calculation

- Calculate average value from sentimental coefficients
- Calculate trendline from known sentimental coefficients and time values $ax^2 + bx + c = y$.

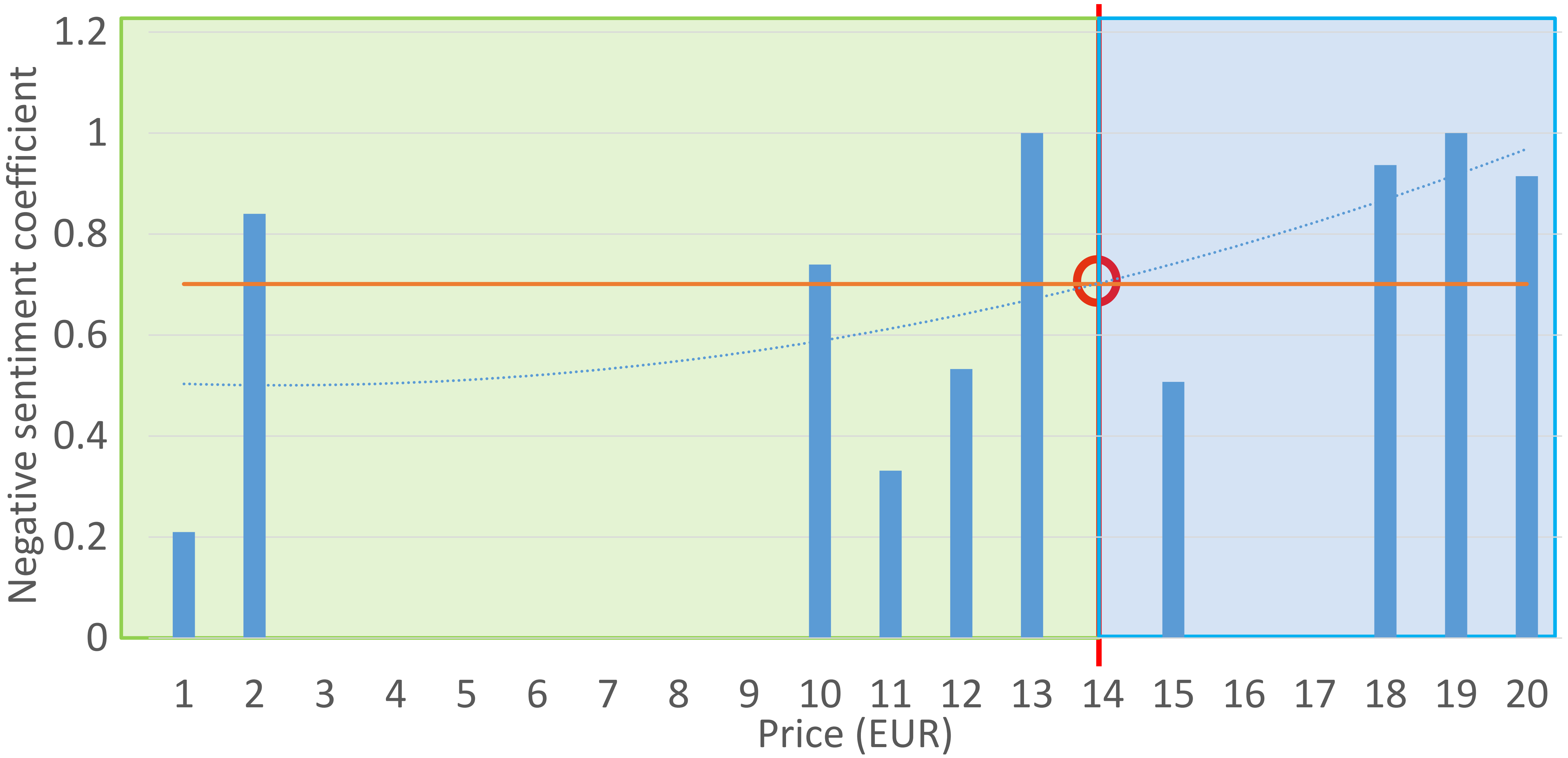


KPI condition specification

- Condition for each KPI theme

KPI theme – *price* less (<)

KPI theme – *insurance compensation* (higher >)



KPI theme	KPI condition	Value interval
ET_1	Higher or Less (> or <)	$X_1 \in [a;b]$
ET_n	Higher or Less (> or <)	$X_n \in [a;b]$

KPI generation

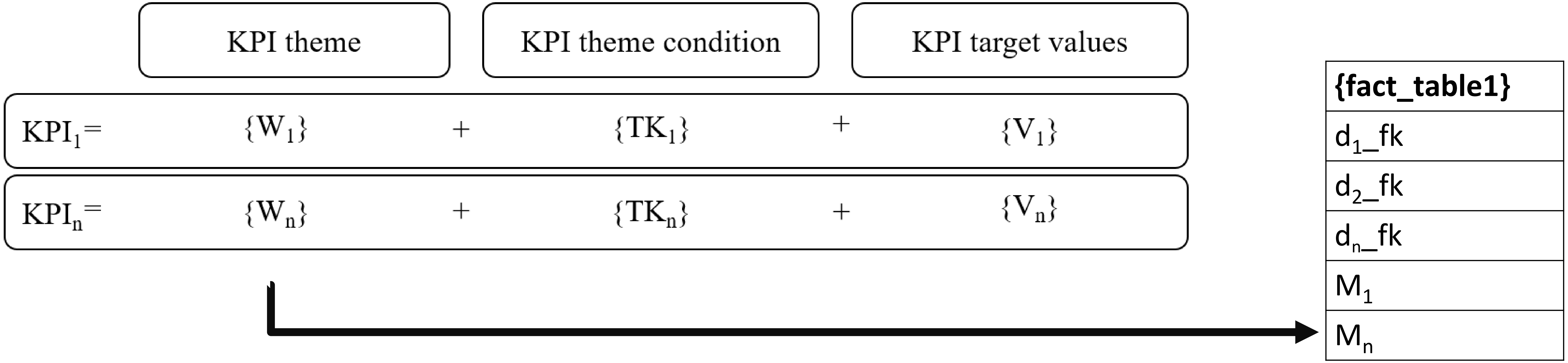
	KPI theme		KPI theme condition		KPI target values
$KPI_1=$	$\{W_1\}$	$+$	$\{TK_1\}$	$+$	$\{V_1\}$
$KPI_n=$	$\{W_n\}$	$+$	$\{TK_n\}$	$+$	$\{V_n\}$
«Price < 14 EUR» =	Price		<		14 EUR

«Price should be less then 14 EUR.»

Data warehouse data model improvements

- Attributes from KPI are integrated into data warehouse data model

Data warehouse multidimensional model fact table



KPI theme | KPI theme condition | KPI target values

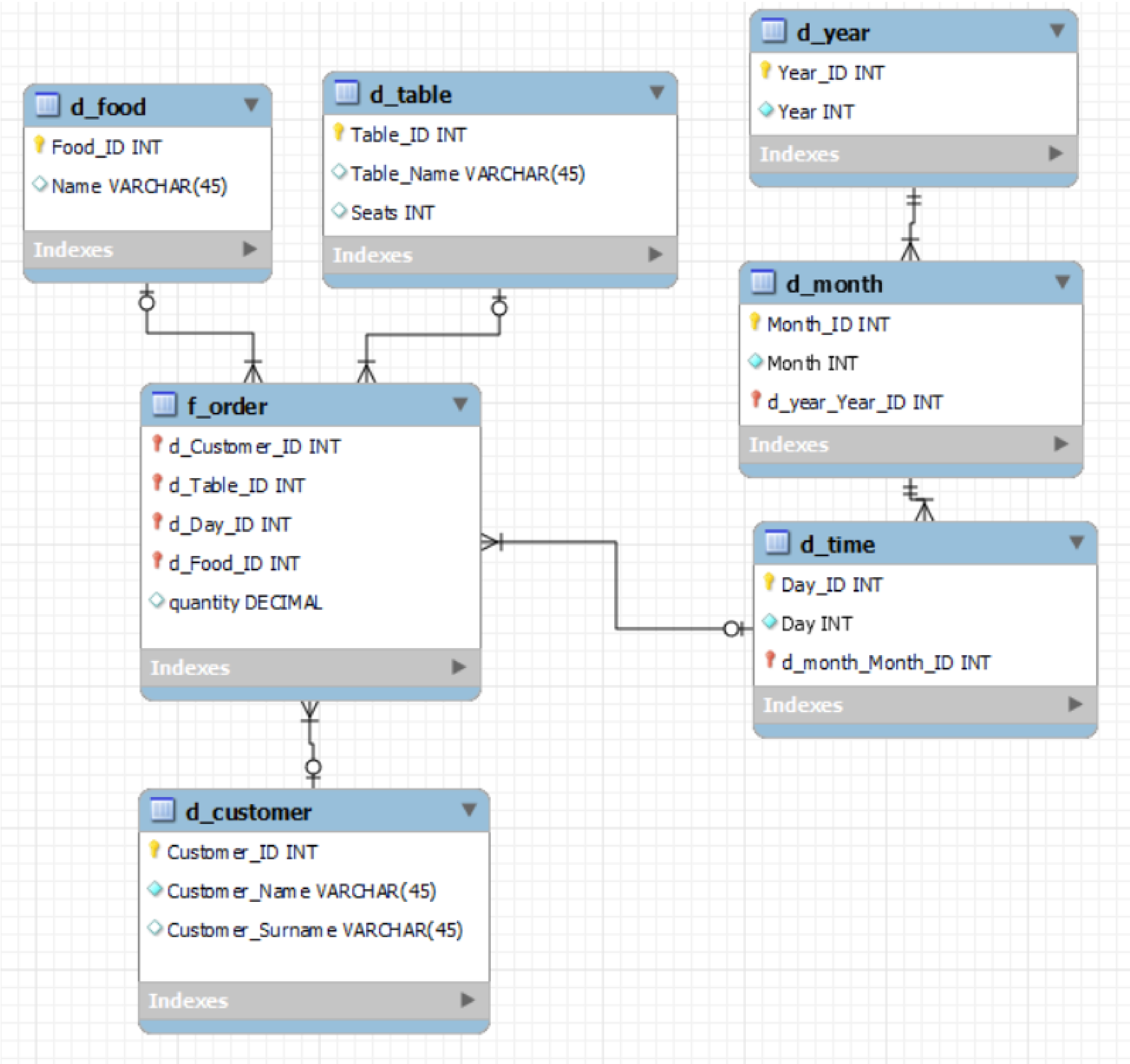
$KPI_1 = \{W_1\} + \{TK_1\} + \{V_1\}$

$KPI_n = \{W_n\} + \{TK_n\} + \{V_n\}$

{fact_table1}
d_1_fk
d_2_fk
d_n_fk
M_1
M_n

UNIVERSITY OF LATVIA

Approbation of the method in an enterprise



d_food
Food_ID INT
Name VARCHAR(45)
Indexes

d_table
Table_ID INT
Table_Name VARCHAR(45)
Seats INT
Indexes

d_year
Year_ID INT
Year INT
Indexes

d_month
Month_ID INT
Month INT
d_year_Year_ID INT
Indexes

f_order
d_Customer_ID INT
d_Table_ID INT
d_Day_ID INT
d_Food_ID INT
quantity DECIMAL
Indexes

d_time
Day_ID INT
Day INT
d_month_Month_ID INT
Indexes

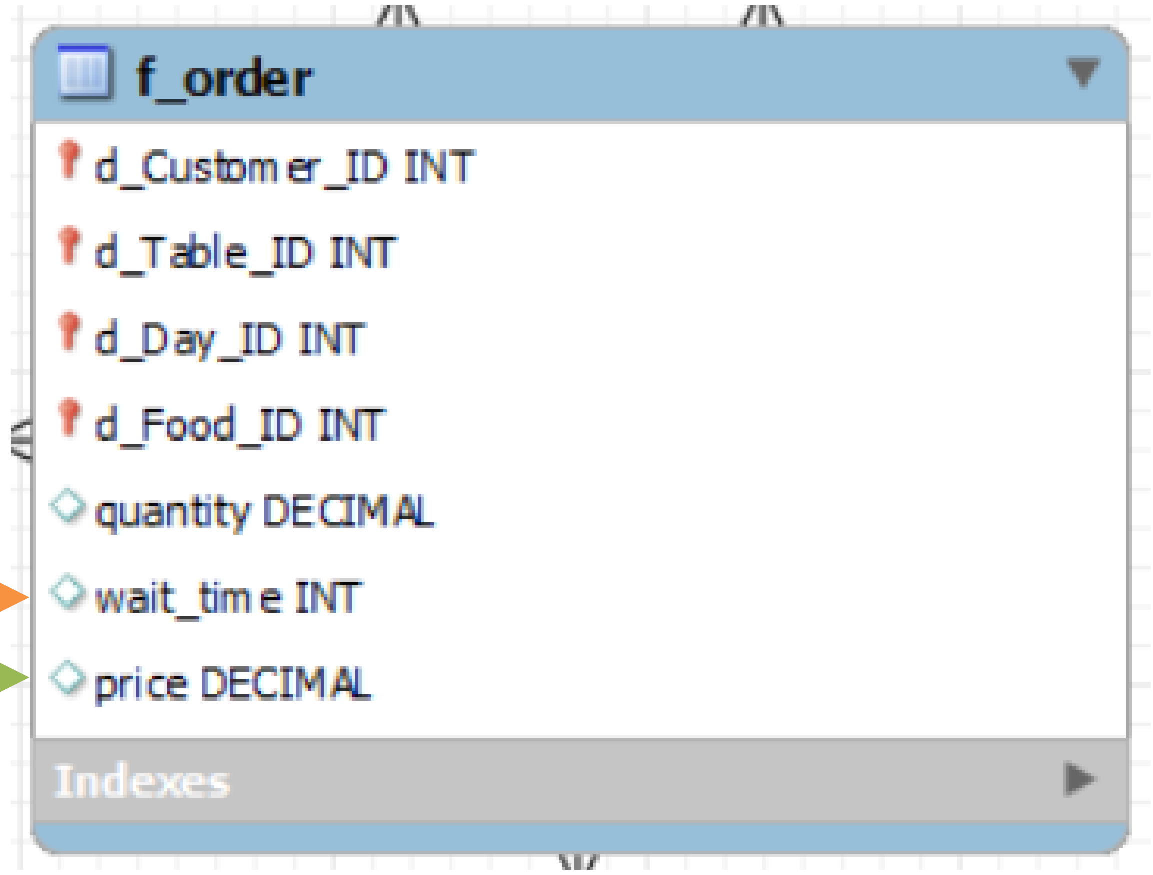
d_customer
Customer_ID INT
Customer_Name VARCHAR(45)
Customer_Surname VARCHAR(45)
Indexes

Approbation of the method in an enterprise - results

Resulting quantitative KPIs

KPI theme	Condition	KPI target value
Price	<	11.5 EUR
Wait time	<	24.9 minutes

Extended multidimensional data model fact table



f_order

- d_Customer_ID INT
- d_Table_ID INT
- d_Day_ID INT
- d_Food_ID INT
- quantity DECIMAL
- wait_time INT
- price DECIMAL

Indexes

Limitations of methods

- In order to perform methods “*Method to define KPIs using unstructured text*” and ”*Data warehouse data model improvements from customer feedback*” organizations must have customer feedback which is used as data source
- In order to generate quantitative KPIs customer feedback must contain numeric values
- Method “Data warehouse data model improvements from customer feedback” is more suitable for organizations with an existing data warehouse

Main results

- Extended existing data warehouse formal requirement metamodel with new classes and developed special tool which implements method

- Developed and practically implemented two new methods :
  - Method to specify new KPI using unstructured data and its GUI
  - Method to specify new quantitative KPI in semi-automated way and data warehouse data model extension by using unstructured data

- During approbation specified new KPIs and improved data warehouse data model

- Published five papers and results are presented in three international conferences

Future improvements

- Location of user feedback creation analysis

- Improved NLP usage:
  - Combine synonyms
  - More detail word dependency analysis
  - Exclude domain specific most used words

- Qualitative/Soft KPIs determination

Publications

- KOZMINA, Natalija; NIEDRITE, Laila; ZEMNICKIS, Janis. Gathering Formalized Information Requirements of a Data Warehouse. In: International Conference on Enterprise Information Systems. SCITEPRESS, 2017. p. 217-224.
- KOZMINA, Natalija; NIEDRITE, Laila; ZEMNICKIS, Janis. Information requirements for big data projects: A review of state-of-the-art approaches. In: Databases and Information Systems: 13th International Baltic Conference, DB&IS 2018, Trakai, Lithuania, July 1-4, 2018, Proceedings 13. Springer International Publishing, 2018. p. 73-89.
- KOZMINA, Natalija; NIEDRITE, Laila; ZEMNICKIS, Janis. Perspectives of Information Requirements Analysis in Big Data Projects. In: Databases and Information Systems X. IOS Press, 2019. p. 109-124.
- ZEMNICKIS, Janis; NIEDRITE, Laila; KOZMINA, Natalija. A Little Bird Told Me: Discovering KPIs from Twitter Data. In: Databases and Information Systems: 14th International Baltic Conference, DB&IS 2020, Tallinn, Estonia, June 16–19, 2020, Proceedings 14. Springer International Publishing, 2020. p. 161-175.
- ZEMNICKIS, Jānis. Data Warehouse Data Model Improvements from Customer Feedback. Baltic Journal of Modern Computing, 2023, 11.3.

Thank you!

UNIVERSITY
OF LATVIA