

RESEARCH PAPER

Data Warehouse Hybrid Modeling Methodology

Viktor László Takács¹, Katalin Bubnó², Gergely Gábor Ráthonyi¹,
Éva Bácsné Bába³ and Róbert Szilágyi¹

¹ Institute of Applied Informatics and Logistics, Faculty of Economics and Business, University of Debrecen, Debrecen, HU

² School of Mathematical and Computational Sciences, University of Debrecen, Debrecen, HU

³ Institute of Rural Development, Tourism and Sports Management, Faculty of Economics and Business, University of Debrecen, Debrecen, HU

Corresponding author: Viktor László Takács (takacs.viktor@econ.unideb.hu)

The classic conceptual modeling around business processes followed by the ‘bus matrix’ methodology of designing the data cubes of data warehouses (Kimball & Ross 2013). For a serious system, such a quantity of management questions and dimensions, the bus matrix results a difficult-to-understand conceptual data model. The subject of automation and conceptual design – to which many individual methods already have been developed – are relevant topics in today’s literature also.

In the 2010s data warehouse projects were realized in Hungarian higher education to inform the decision makers of the universities about their own institutions. As we participated in this project in 2009–2010, we faced that our bus matrix at the end contained about 80–120 indicators with nearly 200 dimensions (dimensional attributes), therefore we worked on the early stenography to formalize the management question.

We provide a kind of ‘business intelligence problem solving thinking’ and a kind of descriptive language that can serve it and present a method which has two novelties compared to formers:

1. It is based on the management questions and its visualization.
2. As a kind of stenography, it is always based on the terminology corresponding to the current problem, so it forms an intermediate language for the data model.

We introduce our method through an example in a popular research area which is activity tracking.

Keywords: Formal methods; Data science; Information Management

1 Introduction

From the 2000’s we had to study a new concept in management sciences. It was ‘Business Intelligence’ (BI). Krauth (2008) summarized the actual information about Business Intelligence and also forecasted the expected development and changes of BI for the next ten years. One of the priorities changes he emphasized was that ‘Technologies providing business intelligence will leave the corporate framework and move on a much wider scale to serve the growing demands of organizations and individuals for accurate, substantial and comprehensible information.’

After ten years we saw that the demand was actually raised in the leaders of small and medium enterprises also for the dissolution of business intelligence, furthermore there are several so-called self-service business intelligence solutions we should use to satisfy this demand. From the 2010s we could see that the basic technology of business informatics the online analytical processing (OLAP) was released in scientific research also, mainly in processing measured data.

In the 2000s business environment – mainly in small and medium enterprises – unfortunately we saw that neither the economists nor the IT staff of these small firms could work with this BI-solutions (softwares). The reason for the problem is not in technical difficulties but in acquiring the right way of thinking to model the function of their own firm, define correctly the information requirement of management (conceptual modeling) and translate it to a logical model.

We participated in a data warehouse projects where we realized that our ‘corporate bus matrix’ contained about 80–120 indicators with nearly 200 dimensions (dimensional attributes), therefore we started to work on the early stenography to formalize the management question (Takács & Bubnó 2012).

2 Related Works

It is an often-mentioned problem today in the literature that there is no standardized or widely agreed method for implementing the conceptual model (Bánné 2012; Macedo & Oliviera 2015; Rizzi 2008). Furthermore, it is a good practice to try to follow the classical design steps of database systems (Halassy 1994) in the design of the data warehouse (conceptual model->logical model->physical model->implementation), but opinions differ in the literature which should be the right order of the steps, furthermore, there is a lot of overlap. Mainly conceptual and logical modeling are often mixed with each other or their borders are blurred, however Halassy clearly defined the levels of database planning, furthermore he proved that these levels must be separated from each other (Halassy 1994).

According to the method presented by us, the conceptual model is nothing more than a set of formalized leadership questions. A management question can already be seen as part of an OLAP data cube, and OLAP cubes can be built from subsets of management questions. The question is to optimize their number and distribution. That is, how many cubes do we have to build and how many questions can be answered. While the former is a matter of financial concern for customers, the latter is about the efficiency of the information system.

As regards data warehouses as information systems, the question of efficiency is most of all the above-mentioned two aspects (cost and amount of information that can be extracted). Di Tria, Lefons, & Tangorra (2017) tested design methodologies and carried out for cost-effectiveness analysis. They have set up their framework and metrics to implement this analysis. The classic approaches of data warehouse design can be sorted into two sets: data driven methods and requirement driven methods. Both have advantages and limitations also (Di Tria, Lefons, & Tangorra 2017). For example, the requirement driven approach leads such multidimensional schemas that usually results in one data cube that answer only one question of the management. The main problem with the multidimensional schemas of the other approach is the big number of the potential questions which cause data lakes that become data swamps. We consider a cuboid as a base of a potential management question. The problem is: what is the minimal (optimal cost) number of cuboids? These problems with both approaches lead to the birth of several so-called hybrid modeling methods to design data warehouses. Di Tria, Lefons, & Tangorra (2017) identified the criteria based on the literature to evaluate these hybrid methods. They used the 4 common and main criteria that is necessary to evaluate a data warehouse designing methodology. These are: Correctness, Completeness, Minimality and Understandability (Halassy 1994). Furthermore, based on the literature they defined metrics to the evaluation of costs and benefits, specially the ‘Metrics for schema quality’ and the ‘Metrics for design effort’. Then they compare six hybrid methods based on this framework:

1. Graph-based Hybrid Multidimensional model (referred to as GrHyMM),
2. UML for Data Warehouse (referred to as UMLDW),
3. Multidimensional Design by Example (referred to as MDBE),
4. Phipps&Davis Methodology (referred to as PDM),
5. Goal-oriented Requirement Analysis for Data Warehouse Design (referred to as GRAnD),
6. Goal/Question/Metric-based Methodology (referred to as GQM).

Based on Di Tria, Lefons, & Tangorra (2017) in **Table 1** we present the steps of the six methodologies above and the features or results of the certain steps, extended with our Visualized Management Question-based Design methodology (referred to as VMQD*).

Our Visualized Management Question-based Design methodology is closest to GQM in the steps, but there are some differences.

1. Requirement Analysis. We collect questions with metrics and dimensionality of the problem also in this phase with the required visualizations from the decision makers via interviews. Then we formalize these specifications with our special structured stenography that is based on the terminology corresponding to the current problem. The output of this step is a set of formalized questions.
2. Deriving minimal granularity. Based on the set of formalized questions we specify the required minimal granularity for every indicator. The output of this step is the set of indicators with minimally detailed dimensions.

Table 1: Steps of the six methodologies.

	GrHyMM	UMLDW	MDBE	PDM	GRaND	GQM	VMQD*
Requirement Analysis	goals, tasks	goals, tasks	queries in SQL	queries in SQL	goals, decisions	goals, questions, metrics	visualized questions, metrics, dimensionality
Minimal Granularity							minimally detailed metrics
Ideal Schema						ideal facts, ideal dimensions	ideal facts, ideal dimensions
Source Analysis	independent, source system schema	independent, CWM	independent	independent	independent	independent, potential schema	potential transactions, attributes, partly dependent, potential schema vs. ideal schema
Integration							
Reconciliation	DB integrity	consistent UML multidimensional schema	DB integrity				
Multidimensional Modeling	facts, attribute tree for facts, remodeling	cubes, dimensions, hierarchies, measures	dimensions and facts from tables	Date dimension and Attribute dimensions for facts MeER	Derived from requirement analysis schemas		MeER
Schema Selection				MeER related to questions			
Manual Refinement			modified automatically generated schema				

3. Deriving ideal schemata. We map the dimensional attributes and values to keys, produce the initial conceptual schemata. The output of this step contains ideal dimensions (keys, attributes and hierarchies) and ideal facts (with dimension keys for join), independently from the sources.
4. Source Analysis. The main question of this step is: What kind of transactions can we get them from? We decompose ideal facts into potential elementary transactional attributes and identify them in the source systems. The output of this step is the derived potential schemata.
5. Integration. Ideal schemata from the requirement analysis are compared with potential star schemata. Match occurs, when the two schemata contain the same fact, and, both have the same dimensionality in the same granularity level. In this step we define required transformations and calculate fact tables and common dimensions with attributes.
6. Multidimensional modeling. We build the cube(s) with dimensions, dimension hierarchies and measures.

We used the Di Tria, Lefons, & Tangorra (2017) notation to visualize the framework of our methodology for better comparison (**Figure 1**).

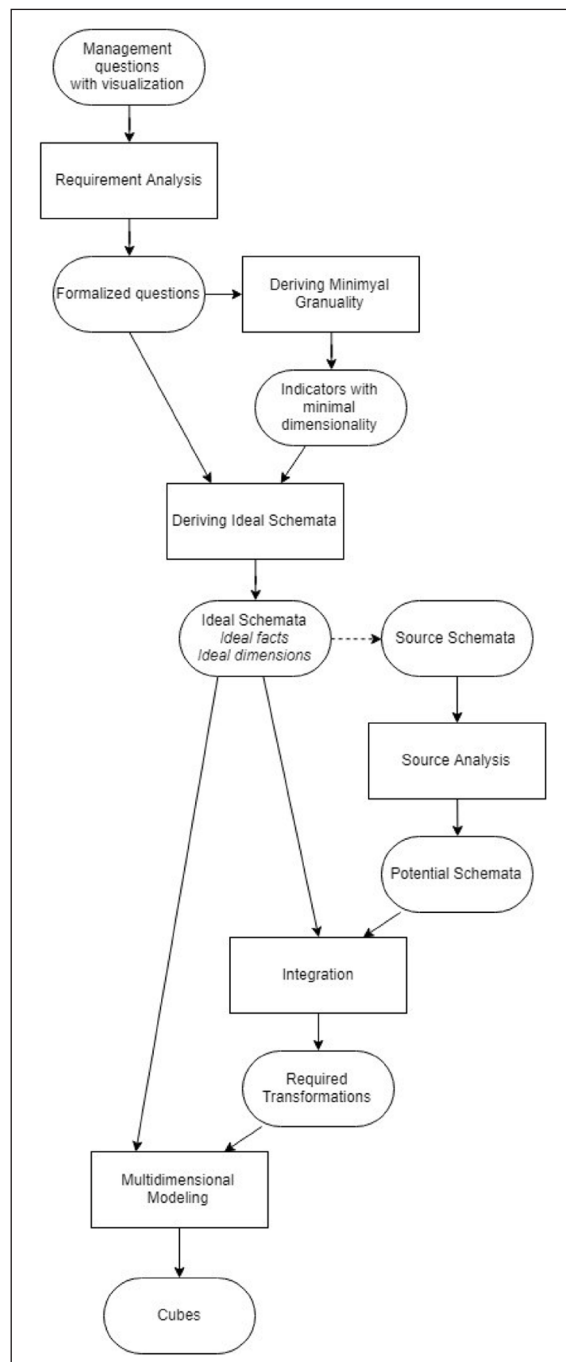


Figure 1: Framework of VMQD.

Every question of the management is a visualization of aggregated indicator(s) detailed with dimensional attribute(s). So, we must formalize these questions of the management.

The formalisation has to cover:

- aggregations (aggregate functions)
i.e. sum, min, max, average, count, median, modus, correlation, etc.;
- indicators with unit
i.e. Price[€], Quantity^{pcs}, Weight^{kg}, etc.;
- visualization
i.e. table, line chart, map, network graphs, population pyramid, etc.;
- attributes
i.e. Gender, Product name, City, Order date, etc.;
- attribute aggregation
i.e. sum, min, max, average, rate, distribution, etc.;
- dimensions
i.e. Person, Product, Geography, Date, etc.;
- dimension hierarchies
i.e. Product->Product category, City->County->Country->Continent

The structure and syntax of the formalization is:

$$Indicator_{\substack{\{units\} \\ \{aggregate\ functions\}}} \left(\begin{matrix} \text{type} \\ [slicer\ attribute] \end{matrix} \right)^{visualization} \left[(\text{attribute values})^{detail} \right]$$

3 Methodology of formalization of conceptual design

In the following paragraph we would like to explain our methodology.

1. In the phase of requirement analysis, we collect and formalize management questions. The question and description are defined in a textual and formal way. The 'manager' is the person for whom the system provides information. As a result, she expects to see a data-visualization report appearing on different dashboards. The reports and dashboards need to be configured according to the requirements of access rights (what kind of level managers can access the statements).

In the requirement analysis, we 'analyse' the management question based on the following considerations in **Table 2**:

What is the indicator? In which aggregation? What is the unit? Which visualization we want to see? In Which detail(s)? Is there a slicer?

Table 2: Management question analysis.

Indicator		the indicator <i>I</i> to be produced with <i>u</i> unit(s) in the upper right index and <i>af</i> aggregate function(s) in the bottom right index,
unit(s)	$I_{\substack{\{u\} \\ \{af\}}}$	
aggregate function(s)		
visualization	$\left(\begin{matrix} vt \\ \left[\begin{matrix} s \\ s \end{matrix} \right] \end{matrix} \right)^v$	the <i>v</i> visualization with the type <i>vt</i> (table, line diagram, bar graph, etc. ...) and optional <i>s</i> slicers (values can be $D_{[a]}$ dimensional attribute, $D_{[v]}$ subset of concrete values, or a $D_{[a]}$ dimensional attribute in the <i>d</i> detail of another <i>I</i> indicator on the same dashboard)
slicer(s)		
detail(s)	$\left[\left[\begin{matrix} D_{\sum\{a\}} \\ D_{\sum\{a\}} \end{matrix} \right] \right]^{(d)}$	<i>d</i> details with $D_{[a]}$ dimensional attribute(s), with optional $\sum\{a\}$ aggregation. <i>d</i> values e.g.: row, column, category, y indicator

$$\text{Formally: } I_{\{af\}^{[u,af]}}^{\{u\}} \left(\left[\begin{array}{c} vt \\ \{s\} \\ \{s\} \end{array} \right] \right)^v \left[\left(\left[\begin{array}{c} D_{\Sigma\{a\}} \\ D_{\Sigma\{a\}} \end{array} \right] \right)^{\{d\}} \right]$$

This formal definition is related to one diagram. Several different visualizations need to arise from the leader on a question related to the requirement specification, so many formal descriptions are made in this section.

2. Deriving the minimum granularity for each indicator. Determine the indicators in the management question and determine the required minimum granularity (dimension and key), where:
 - $I^{\{u\}}$ the indicator I with u unit(s) in the upper right index
 - $D_{\{dk\}}$ dimension D with dimension key dk in the lower right index

$$\text{Formally: } I^{\{u\}} \left(D_{\{dk\}} \left[D_{\{dk\}} \right] \right)$$

3. When we derive an ideal schemata, we produce fact tables for the indicators specified in the minimum granularity and optimizing the number of fact sheets. The properties are grouped into dimensions and hierarchies. Each dimension must have a dimension key (see relational data model), where:
 - $I^{\{u\}}$ the indicator I with u unit(s) in the upper right index
 - $D_{\{dk\}}$ dimension D with dimension key dk in the lower right index
 - $D_{\{a\}}$ dimension D with attribute a in the lower right index
 - $D_{\{i\}}$ dimension D with indicator i in the lower right index
 - $D_{\{dhk\}}$ dimension D with dimension hierarchy key dhk in the lower right index, when hierarchy levels are stored in separate tables

Fact tables formally: $I^{\{u\}} \left(D_{\{dk\}} \left[D_{\{dk\}} \right] \right)$, or $\left(\left[\begin{array}{c} I^{\{u\}} \\ I^{\{u\}} \end{array} \right] \right) \left(D_{\{dk\}} \left[D_{\{dk\}} \right] \right)$ depending on whether the given fact table contains one or more indicators.

$$\text{Dimension table formally: } D \left(D_{\{dk\}} \left[D_{\{a\}} \right] \left[D_{\{i\}} \right] \left[D_{\{dhk\}} \right] \right)$$

During optimization we can notice one of the following cases in **Table 3**:

- Different indicators with the same dimensionality and granularity can be grouped into one fact table.
- Different indicators could be similar when they have the same dimensionality but diverging granularity, so we examine whether they can be produced from one another.
- Dimensionality of one indicator is a proper subset of the dimensionality of the other indicator, so the other indicator dimensionality will be the minimum required.

Table 3: Optimizations' notations.

$\left(\begin{array}{c} I_1 \\ I_2 \end{array} \right) \left(D_{\{dk\}} \right) \equiv I_1 \left(D_{\{dk\}} \right) \times I_2 \left(D_{\{dk\}} \right)$	Combining indicators I_1 and I_2 with the same dimensionality. We create the Descartes multiplier of the two indicators.
$I \left(\left[all, D_{\{dhk\}} \right], D_{\{dk\}} \right) \equiv_{af} I \left(\sum_{D_{\{dk\}}} \left[all, D_{\{dhk\}} \right] \right)$	The value of the indicator I can be obtained by summing through dimension D (roll up) with the aggregate function in the lower left index of I . Calculating the aggregation from $D_{\{dk\}}$ at the bottom of the Summa symbol to the level at the top of the Summa sign (all or $D_{\{dhk\}}$ hierarchy level, leaving the original key. This is referred to as $D_{\{dk\}}$.
$I_1 \left(A_{\{dk\}} \right) \subset I_2 \left(A_{\{dk\}}, B_{\{dk\}} \right)$ $\left(\begin{array}{c} I_1 \left(A_{\{dk\}} \right) \\ I_2 \left(A_{\{dk\}}, B_{\{dk\}} \right) \end{array} \right) \equiv \left(\begin{array}{c} I_1 \\ I_2 \end{array} \right) \left(A_{\{dk\}}, B_{\{dk\}} \right)$	A and B are dimensions of indicators I_1 and I_2 and I_1 is proper subset of I_2 .

4. We will evaluate the available data during the source analysis. Source system can be transaction-oriented or analysis-oriented.
- The Transaction-Oriented Source System (OLTP) is characterized by a relational data model

$$E\left(E_{\{pk\}}[E_{\{a\}}][E_{\{fk\}}]\right) \text{ and } R\left(E_{\{pk\}}[E_{\{pk\}}], R_{\{a\}}[R_{\{a\}}]\right)$$

where:

- $E()$ E entity in the source system,
- $E_{\{pk\}}$ pk primary key of E entity
- $E_{\{a\}}$ a attribute of E entity
- $E_{\{fk\}}$ fk foreign key of E entity
- $R()$ R relation, transaction between different entities, with the relational attributes
- $R_{\{a\}}$ a attribute of R relation

- In the case of an Analysis-Oriented Source System (OLAP), the multidimensional relational model is typical.

Fact tables formally: $I^{(u)}(D_{\{dk\}}[D_{\{dk\}}])$, or $\left(\begin{matrix} I^{(u)} \\ [I^{(u)}] \end{matrix}\right)(D_{\{dk\}}[D_{\{dk\}}])$ depending on whether the given fact table contains one or more indicators.

Dimension table formally: $D(D_{\{dk\}}[D_{\{a\}}][D_{\{i\}}][D_{\{dhk\}}])$, where:

- $I^{(u)}$ the indicator I with u unit(s) in the upper right index
- $D_{\{dk\}}$ dimension D with dimension key dk in the lower right index
- $D_{\{a\}}$ dimension D with attribute a in the lower right index
- $D_{\{i\}}$ dimension D with indicator i in the lower right index
- $D_{\{dhk\}}$ dimension D with dimension hierarchy key dhk in the lower right index, when hierarchy levels are stored in separate tables

5. During integration, we coordinate the ideal data model with the available source system data. Describe the requirements that include attributes, dimension keys, and additional indicators to use. We design the data loading (ETL/ELT) process in **Table 4**.

In the data loading process, extracting, transforming, and loading only those keys, attributes, and indicators from the source system to the data warehouse that we need.

During the transformation of data, two types of data loading processes can be discussed:

- During ETL (Extract, Transform, Load) process the various data manipulations carried out after extracting the data from the source system are typically in an intermediate system (Staging Area) and then transferred to the data warehouse. Intermediate use of complex multi-step data manipulations (transform) is needed, for simpler systems it is worth thinking about using it as it can serve more data warehouses with intermediate data than a commonly used general date dimension, geographical dimension, organizational dimension (enterprise data lake).

The transformations can be simple, then the indicators are loaded from one source system relation, and the dimensions are also derived from a single source system.

$$R\left(E_{\{pk\}}[E_{\{pk\}}], R_{\{a\}}[R_{\{a\}}]\right) \xRightarrow{etl} \left(\begin{matrix} I^{(u)} \\ [I^{(u)}] \end{matrix}\right)(D_{\{dk\}}[D_{\{dk\}}])$$

$$E\left(E_{\{pk\}}[E_{\{a\}}][E_{\{fk\}}]\right) \xRightarrow{etl} D\left(D_{\{dk\}}[D_{\{a\}}][D_{\{i\}}][D_{\{dhk\}}]\right)$$

The transformations may be more complicated when the values of the indicators are loaded from several relations' attributes or from several different source systems. This also occurs with the indicators of dimension.

- During ELT (Extract, Load, Transform) process the various data manipulations take place after the data is extracted from the source system and loaded into the data warehouse.

Table 4: Data loadings' transformation notations.

$I\left(\left[all, D_{\{dhk\}}, D_{\{a\}}\right], \mathcal{D}_{\{dhk\}}\right) = \underset{af}{I}\left(\sum_{D_{\{dhk\}}} \left[all, D_{\{dhk\}}, D_{\{a\}}\right]\right)$	<p>The value of the indicator I can be obtained by summing through D dimension (roll up) with the aggregate function in the lower left index of I. This is an aggregation is from $D_{\{dhk\}}$ at the bottom of the Summa symbol to the level at the top of the Summa Sign (all or $D_{\{dhk\}}$ hierarchy level, leaving the original key. This is referred to as $\mathcal{D}_{\{dhk\}}$.</p>
$D\left(D_{\{dhk\}}\left[\left[D_{\{a\}}\right]\right], D_{\{i\}}\right) = D\left(\sum_{D_{\{dhk\}}} \left[\left[D_{\{a\}}\right]\right], \underset{af}{D}_{\{i\}}\right)$	<p>Deduplicate the values of D dimensions' $D_{\{dhk\}}$ key. Summarize the indicator with the af aggregate function in the lower left index, while leaving the first element of attribute values.</p>
$I\left(D_{\{dhk\}}\right) = I\left(\right) \times D_{\{dhk\}}$	<p>Expand the dimensionality of indicator I. The Descartes multiplier of the original indicator with the dimension to be expanded.</p>
$I_{D_{\{a\}}}(\) = I\left(\circlearrowleft D_{\{a\}}\right)$	<p>Pivoting I indicator values through $D_{\{a\}}$ dimensional attribute. We create several new indicators corresponding to the occurrence values of the attribute.</p>
$\left(\begin{matrix} I_1 \\ I_2 \end{matrix}\right)\left(D_{\{dhk\}}\right) = I_1\left(D_{\{dhk\}}\right) \times I_2\left(D_{\{dhk\}}\right)$	<p>Combining I_1, I_2 indicators with the same dimensionality. We create the Descartes multiplier of the two indicators.</p>
$V^{\{u_1, u_2\}}\left(D_{\{dhk\}}, \mathcal{A}_{\{h_1, h_2\}}\right) = \left(\circlearrowleft \begin{matrix} I_1^{\{u_1\}} \\ I_2^{\{u_2\}} \end{matrix}\right)\left(D_{\{dhk\}}\right)$	<p>Unpivoting I_1, I_2 indicators with the same dimensionality into V indicator values and \mathcal{A} attribute set with the indicators' name</p>
$I = \sum_{D_{\{a\}}}^{all} I_{D_{\{a\}}}$	<p>The sum of pivoted $I_{D_{\{a\}}}$ indicator values along the occurrence values of $D_{\{a\}}$ attribute.</p>

When we formalize a business question, the method is pure requirement-driven. We do not bother with data-driven approach at this stage. We bother with it after we have collected and formalized and modelled all the questions, and the required data model is generated, then we examine the questions through the data-driven approach also. At this point we examine how the formalized elements of the questions are stored in the source systems. If we can produce them in the required granularity (or more detailed) then we optimize them (for example with the use of some kind of aggregation function) to the ETL/ELT (Extract-Transform-Load/Extract-Load-Transform) processes.

4 Research data processing example (Collecting and processing fitness tracker data)

The relationship between physical inactivity and some chronic health conditions is a widely researched area but further efforts are needed to assist people to adopt healthier lifestyles (Lee et al. 2012). Using wearable activity trackers can be a promising opportunity for individuals to improve lifestyle behaviour (Maher et al. 2017). There are several studies in this area, mainly from the lifestyle behaviour and health approach (Henriksen et al. 2018; Kaewkannate & Kim 2016; Karapanos et al. 2016). Our research does not examine activity trackers from the aspect of health, we want to present the way that we can process data with OLAP technology we collected with a very simple device. If an end user who wants to know his own activity by using such an activity tracker, usually he downloads a software that processes his data every day and informs him. But if we plan a wide research where we want to collect several persons' data, and we want to recognize trends and patterns in the behaviour of the society, it could be a useful approach if we plan a data cube with OLAP approach. With our hybrid design methodology, we get metadata from the dimensions and attributes, so if we extract our data into a dataset, we get the formal description of the structure of our dataset also, in order to share and compare it to other similar researches.

Our research investigated the physical activity of university students using fitness tracker. Participants in the pilot test had to meet several criteria. Participants had to wear the device with normal living conditions for 90 consecutive days, which simulated the normal living conditions of most students. An important element of the long-term pilot test is that it can represent the full range of normal people's activities in a real environment. Each participant was informed on the most important information about the device and the potential for managing possible sources of error. The battery of the bracelet was recharged by the users every 20 days, depending on the use, whereby the data was collected at the same time. The data was sent by the users for one week on a daily basis and then at the charges mentioned above. This level of data supply has served to reduce the potential loss of data. We informed the students about the study and all the participants provided informed consent in compliance with the principles of the Declaration of Helsinki (WMA 2013) and the new GDPR (EP 2016). The study was approved by the Regional Ethics Board (DE RKEB/IKÉB: 4843-2017) at the Clinical Center of the University of Debrecen.

Collected bracelets' data are processed using OLAP technology. We use the following hybrid design methodology and formal descriptive techniques to design, implement, and document the operations related to the information system (Research Data Warehouse) that we produce.

4.1 Requirement analysis

Question1 formally in Table 5: The students' daily activity by daily steps intensity categories in March:

It shows how many days were completed by the students in March by daily step categories.

The 'daily step categories' naturally require a detailed discussion during the requirement analysis and at the same time predicts a clustering task in the integration phase.

Table 5: Question1 analysis.

Indicator	how many days completed (activity)		
unit(s)	day	$I_{\{af,af\}}^{\{d,af\}}$	$Activity^{\{day\}}$
aggregate function(s)	how many (sum)		
visualization	table	$\left(\begin{matrix} vt \\ \left[\begin{matrix} s \\ s \end{matrix} \right] \end{matrix} \right)^v$	$\left(\begin{matrix} table \\ \mathbf{D}_{\{March\}} \end{matrix} \right)^v$
slicer(s)	March		
	student	$\left[\left(\begin{matrix} \mathbf{D}_{\{a\}} \end{matrix} \right)^{\{d\}} \right]$	$(\mathbf{P}_{\{stud\}})^{row}$
detail(s)	daily step category	$\left[\left(\begin{matrix} \mathbf{D}_{\{a\}} \end{matrix} \right)^{\{d\}} \right]$	$(\mathbf{I}_{\{dsc\}})^{col}$

$$Activity^{\{day\}} \left(\begin{matrix} table \\ \mathbf{D}_{\{March\}} \end{matrix} \right)^v (\mathbf{P}_{\{nid\}})^{row} (\mathbf{I}_{\{dsc\}})^{col}$$

Question2 formally in Table 6: Students' average daily activity in March by category and gender:

It shows the students' averagely completed days in March by daily step categories and gender.

The term 'students' and 'daily' refers to the maximal details of the data average that we can deal with in the minimal granularity, optimal data model or in the integration phase.

Table 6: Question2 analysis.

Indicator	averagely completed days		
unit(s)	day	$I_{\{af[af]\}}^{\{u[af]\}}$	$Activity_{average}^{\{day\}}$
aggregate function(s)	average		
visualization	table	$\left(\begin{matrix} vt \\ \left[\begin{matrix} s \\ s \end{matrix} \right] \end{matrix} \right)^v$	$\left(\begin{matrix} table \\ \mathbf{D}_{\{March\}} \end{matrix} \right)^v$
slicer(s)	March		
detail(s)	gender	$\left[\begin{matrix} \left(\mathbf{D}_{\{a\}} \right)^{\{f\}} \\ \left[\mathbf{D}_{\{a\}} \right] \end{matrix} \right]$	$\left(P_{\{gender\}} \right)^{row}$
	daily step category	$\left[\begin{matrix} \left(\mathbf{D}_{\{a\}} \right)^{\{f\}} \\ \left[\mathbf{D}_{\{a\}} \right] \end{matrix} \right]$	$\left(I_{\{dsc\}} \right)^{col}$

$$Activity_{average}^{\{day\}} \left(\begin{matrix} table \\ \mathbf{D}_{\{March\}} \end{matrix} \right)^v \left(P_{\{gender\}} \right)^{row} \left(I_{\{dsc\}} \right)^{col}$$

Question3 formally in Table 7: Average daily steps of men, women and all by the day of the week in March:

It shows a comparison of mens', womens' and combined average daily number of steps in March, in the days of the week.

Table 7: Question3 analysis.

Indicator	Daily steps		
unit(s)	steps	$I_{\{af[af]\}}^{\{u[af]\}}$	$DailySteps_{average}^{\{steps\}}$
aggregate function(s)	average		
visualization	radar chart	$\left(\begin{matrix} vt \\ \left[\begin{matrix} s \\ s \end{matrix} \right] \end{matrix} \right)^v$	$\left(\begin{matrix} radar\ chart \\ \mathbf{D}_{\{March\}} \end{matrix} \right)^v$
slicer(s)	March		
detail(s)	day of the week	$\left[\begin{matrix} \left(\mathbf{D}_{\{a\}} \right)^{\{d\}} \\ \left[\mathbf{D}_{\{a\}} \right] \end{matrix} \right]$	$\left(\mathbf{D}_{\{Dow\}} \right)^{cat}$
	men, women, all	$\left[\begin{matrix} \left(\mathbf{D}_{\{a\}} \right)^{\{d\}} \\ \left[\mathbf{D}_{\{a\}} \right] \end{matrix} \right]$	$\left(P_{\Sigma\{gender\}} \right)^y$

$$DailySteps_{average}^{\{steps\}} \left(\begin{matrix} radar\ chart \\ \mathbf{D}_{\{March\}} \end{matrix} \right)^v \left(\mathbf{D}_{\{weekday\}} \right)^{cat} \left(P_{\Sigma\{gender\}} \right)^y$$

4.2 Minimum granularity for each indicator

We define attributes for values and keys for dimensional attributes in the questions.

$$Activity_{average}^{\{day\}} \left(\begin{matrix} table \\ \mathbf{D}_{\{March\}} \end{matrix} \right)^v \left(P_{\{stud\}} \right)^{row} \left(I_{\{dsc\}} \right)^{col\ min} \Rightarrow Activity^{\{day\}} \left(P_{\{PK\}}, \mathbf{D}_{\{MK\}}, I_{\{IK\}} \right)$$

March is month value of Date dimension ($\mathbf{D}_{\{\text{March}\}}$) so the related dimension key should be MonthKey ($\mathbf{D}_{\{\text{MK}\}}$)

Neptun ID is an attribute of Person dimension ($\mathbf{P}_{\{\text{stud}\}}$) so the related dimension key should be PersonKey ($\mathbf{P}_{\{\text{PK}\}}$).

Daily step category is an attribute of activity Intensity dimension ($\mathbf{I}_{\{\text{dsc}\}}$), so the dimension key should be IntensityKey ($\mathbf{I}_{\{\text{IK}\}}$).

$$\mathbf{Activity}_{\text{average}}^{\{\text{day}\}} \left(\begin{array}{c} \text{table} \\ \mathbf{D}_{\{\text{March}\}} \end{array} \right)^v \left(\mathbf{P}_{\{\text{gender}\}} \right)^{\text{row}} \left(\mathbf{I}_{\{\text{dsc}\}} \right)^{\text{col}} \stackrel{\text{min}}{\Rightarrow} \mathbf{Activity}^{\{\text{day}\}} \left(\mathbf{P}_{\{\text{GK}\}}, \mathbf{D}_{\{\text{MK}\}}, \mathbf{I}_{\{\text{IK}\}} \right)$$

March is month value of Date dimension ($\mathbf{D}_{\{\text{March}\}}$) so the related dimension key should be MonthKey ($\mathbf{D}_{\{\text{MK}\}}$)

Gender is an attribute of Person dimension ($\mathbf{P}_{\{\text{gender}\}}$) so the possible dimension key should be PersonKey ($\mathbf{P}_{\{\text{PK}\}}$) or GenderKey ($\mathbf{P}_{\{\text{GK}\}}$), the minimum granularity is GenderKey ($\mathbf{P}_{\{\text{GK}\}}$).

Daily step category is an attribute of activity Intensity dimension ($\mathbf{I}_{\{\text{dsc}\}}$), so the dimension key should be IntensityKey ($\mathbf{I}_{\{\text{IK}\}}$).

$$\mathbf{DailySteps}_{\text{average}}^{\{\text{steps}\}} \left(\begin{array}{c} \text{radar chart} \\ \mathbf{D}_{\{\text{March}\}} \end{array} \right)^v \left(\mathbf{D}_{\{\text{weekday}\}} \right)^{\text{cat}} \left(\mathbf{P}_{\sum\{\text{gender}\}} \right)^y \stackrel{\text{min}}{\Rightarrow} \mathbf{DailySteps}^{\{\text{steps}\}} \left(\mathbf{P}_{\{\text{GK}\}}, \mathbf{D}_{\{\text{DK}\}} \right)$$

March is month value of Date dimension ($\mathbf{D}_{\{\text{March}\}}$) and weekday is an attribute of Date dimension ($\mathbf{D}_{\{\text{weekday}\}}$) so the related dimension key should be MonthKey ($\mathbf{D}_{\{\text{MK}\}}$) and DayofWeek ($\mathbf{D}_{\{\text{DoW}\}}$), so the common dimension key for both is DateKey ($\mathbf{D}_{\{\text{DK}\}}$).

Gender is an attribute of Person dimension ($\mathbf{P}_{\{\text{gender}\}}$) so the possible dimension key should be PersonKey ($\mathbf{P}_{\{\text{PK}\}}$) or GenderKey ($\mathbf{P}_{\{\text{GK}\}}$), the minimum granularity is GenderKey ($\mathbf{P}_{\{\text{GK}\}}$).

The following indicators with the minimal required granularity should be the base to answers each question we have:

$$\mathbf{Activity}^{\{\text{day}\}} \left(\mathbf{P}_{\{\text{PK}\}}, \mathbf{D}_{\{\text{MK}\}}, \mathbf{I}_{\{\text{IK}\}} \right)$$

$$\mathbf{Activity}^{\{\text{day}\}} \left(\mathbf{P}_{\{\text{GK}\}}, \mathbf{D}_{\{\text{MK}\}}, \mathbf{I}_{\{\text{IK}\}} \right)$$

$$\mathbf{DailySteps}^{\{\text{steps}\}} \left(\mathbf{P}_{\{\text{GK}\}}, \mathbf{D}_{\{\text{DK}\}} \right)$$

4.3 Ideal schemata for OLAP system

In this step we determine which indicators can be stored in a common fact table.

During optimization, we can see which indicators are similar and can be produced from one another. In this case activity indicator detailed by GenderKey (GK) can be generated from the activity indicator detailed by PersonKey (PK). This means that we have already managed to optimize the number of indicators we are building.

$$\mathbf{Activity}^{\{\text{day}\}} \left(\mathbf{P}_{\{\text{GK}\}}, \mathbf{D}_{\{\text{MK}\}}, \mathbf{I}_{\{\text{IK}\}} \right) \equiv \mathbf{Activity}^{\{\text{day}\}} \left(\sum_{\mathbf{P}_{\{\text{PK}\}}}^{\mathbf{P}_{\{\text{GK}\}}}, \mathbf{D}_{\{\text{MK}\}}, \mathbf{I}_{\{\text{IK}\}} \right)$$

$$\mathbf{Activity}^{\{\text{day}\}} \left(\sum_{\mathbf{P}_{\{\text{PK}\}}}^{\mathbf{P}_{\{\text{GK}\}}}, \mathbf{D}_{\{\text{MK}\}}, \mathbf{I}_{\{\text{IK}\}} \right) \stackrel{\text{min}}{\Leftarrow} \mathbf{Activity}^{\{\text{day}\}} \left(\mathbf{P}_{\{\text{PK}\}}, \mathbf{D}_{\{\text{MK}\}}, \mathbf{I}_{\{\text{IK}\}} \right)$$

The average DailySteps indicator must break down into gender variants and the aggregated.

$$\mathbf{DailySteps}_{\text{average}}^{\{\text{steps}\}} \left(\sum_{P_{\{GK\}}}^{\text{all}}, \mathbf{D}_{\{DK\}} \right) \equiv \mathbf{AverageDailySteps}^{\{\text{steps}\}} \left(\mathbf{D}_{\{DK\}} \right)$$

$$\mathbf{DailySteps}_{\text{average}}^{\{\text{steps}\}} \left(\bigcirc \sum_{P_{\{GK\}}}^{P_{\{\text{gender}\}}}, \mathbf{D}_{\{DK\}} \right) \cap P_{\{\text{Male}\}} \equiv \mathbf{AverageDailySteps}_{\text{Male}}^{\{\text{steps}\}} \left(\mathbf{D}_{\{DK\}} \right)$$

$$\mathbf{DailySteps}_{\text{average}}^{\{\text{steps}\}} \left(\bigcirc \sum_{P_{\{GK\}}}^{P_{\{\text{gender}\}}}, \mathbf{D}_{\{DK\}} \right) \cap P_{\{\text{Female}\}} \equiv \mathbf{AverageDailySteps}_{\text{Female}}^{\{\text{steps}\}} \left(\mathbf{D}_{\{DK\}} \right)$$

Next step is to place indicators into fact tables. Indicators with the same dimensionality and granularity are placed into a common fact table. In this case these are the daily activity and the daily steps fact tables.

ftDailyActivity: Daily activity fact table

$$\mathbf{Activity}^{\{\text{day}\}} \left(P_{\{PK\}}, \mathbf{D}_{\{MK\}}, \mathbf{I}_{\{IK\}} \right)$$

ftDailySteps: Daily steps fact table

$$\left(\begin{array}{l} \mathbf{AverageDailySteps}^{\{\text{steps}\}} \\ \mathbf{AverageDailySteps}_{\text{Male}}^{\{\text{steps}\}} \\ \mathbf{AverageDailySteps}_{\text{Female}}^{\{\text{steps}\}} \end{array} \right) \left(\mathbf{D}_{\{DK\}} \right) \equiv \mathbf{AverageDailySteps}_{P_{\{\Sigma, \text{gender}\}}}^{\{\text{steps}\}} \left(\mathbf{D}_{\{DK\}} \right)$$

Dimensional attributes and values in questions and dimensional keys in the minimal and ideal data model must be organized into dimensions. In this step we specify the Dimension->Key-Attribute-Indicator->Value structures, also the required dimension hierarchies with hierarchy-keys. We have three dimensions (Person, Date, Intensity) in our ideal data model in the following structure:

P: Person (dimPerson)

- PersonKey ($P_{\{PK\}}$): HASH of students' identifier
- Student ($P_{\{\text{stud}\}}$): students' identifier (eliminated, because of GDPR)
- GenderKey ($P_{\{GK\}}$): unique identifier, values {1, 2} (eliminated, when we unfold the dimension hierarchy)
- Gender ($P_{\{\text{gender}\}}$): gender of students, values {Male, Female}

D: Date (dimDate)

- DateKey ($D_{\{DK\}}$): **year**, **month**, **day** serial number with leading zero, composition {yyyy}{mm}{dd}
- Day of Week ($D_{\{\text{Dow}\}}$): unique identifier, serial number of weekdays {1..7} (eliminated, when we partly unfold the dimension hierarchy)
- Weekday ($D_{\{\text{weekday}\}}$): {1 – Monday, 2 – Tuesday, ..., 7 – Sunday}
- MonthKey ($D_{\{MK\}}$): **year**, **month** serial number with leading zero, composition {yyyy}{mm}
- **DM:** DateMonth (dimDateMonth)
 - MonthKey ($DM_{\{MK\}}$): **year**, **month** serial number with leading zero, composition {yyyy}{mm}
 - Month ($DM_{\{\text{month}\}}$): month name, values {January, February, ..., December}

I: Intensity (dimIntensity): Categories of daily activity of students.

- IntensityKey ($I_{\{IK\}}$): {0..5}
- Daily Step Category ($I_{\{\text{dsc}\}}$):
 - 0 – Basal activity
 - 1 – Limited activity

- 2 – Low activity
- 3 – Somewhat active
- 4 – Active
- 5 – Highly active

4.4 Source analysis

In this phase we discover the data of the source systems driven by the facts and dimensions specified in the ideal data model.

4.4.1 Activity tracker data

$\mathcal{S}_{iOS}(\mathcal{S}_{\{steps\}}, \mathcal{S}_{\{timestamp\}}, \mathcal{S}_{\{date\}}, \mathcal{S}_{\{10min\}})$ Emails sent via email from iPhone, the name of the file contains the student's Neptun identifier and the date of submission in the $\mathcal{S}_{\{NID\}}-\mathcal{S}_{\{sd\}}$ structure.

$\mathcal{S}_A(\mathcal{S}_{\{cumulative\ steps\}}, \mathcal{S}_{\{timestamp\}}, \mathcal{S}_{\{date\}}, \mathcal{S}_{\{min\}})$ Emails sent via email from Android phone, the name of the file contains the student's Neptun identifier and the date of submission in the $\mathcal{S}_{\{NID\}}-\mathcal{S}_{\{sd\}}$ structure.

\mathcal{S} : Steps (noted as rSteps) with the following attributes and values:

- Date ($\mathcal{S}_{\{date\}}$): {year}.{month}.{day}
- DateKey ($\mathcal{S}_{\{DK\}}$): {year}{month}{day}
- 10mins ($\mathcal{S}_{\{10mins\}}$): {hh}:{m0}:{00}
- minute ($\mathcal{S}_{\{min\}}$): {hh}:{mm}:{00}
- TimeKey ($\mathcal{S}_{\{TK\}}$): $\mathcal{S}_{\{10min\}}$ mapping to [0..143] integers closed interval
- timestamp ($\mathcal{S}_{\{timestamp\}}$): in seconds, the number of seconds passed since 1900.01.00 0:00:00
- StepSum_{iOS} ($\mathcal{S}_{\{steps\}}$): number of steps taken in 10 minutes
- StepSum_A ($\mathcal{S}_{\{cumulative\ steps\}}$): the number of daily steps taken to a given time
- StepType ($\mathcal{S}_{\{st\}}$): {raw, normalized}
- 10minNS ($\mathcal{S}_{\{10mNS\}}$): 10-minute normalized step data
- Neptun identifier ($\mathcal{S}_{\{NID\}}$): the NEPTUN identifier of the student submitting the data
- Person key ($\mathcal{S}_{\{PK\}}$): the hashed NEPTUN identifier of the student submitting the data
- Submission date ($\mathcal{S}_{\{sd\}}$): date of data submission in {year}{month}{day} structure
- ETL date ($\mathcal{S}_{\{etld\}}$): the date key of the ETL process in {year}{month}{day} structure
- From 'source systems' the data is loaded into an intermediate storage.
- We supplement each standalone file with the data in its name and the step type dependent on the mobile operating system.

$$\mathcal{S}_{iOS}(\mathcal{S}_{\{steps\}}, \mathcal{S}_{\{timestamp\}}, \mathcal{S}_{\{date\}}, \mathcal{S}_{\{10min\}}) \times (\mathcal{S}_{\{NID\}}, \mathcal{S}_{\{sd\}}, \mathcal{S}_{\{etld\}}) \times \mathcal{S}_{\{normalized\}}$$

$$\mathcal{S}_A(\mathcal{S}_{\{cumulative\ steps\}}, \mathcal{S}_{\{timestamp\}}, \mathcal{S}_{\{date\}}, \mathcal{S}_{\{min\}}) \times (\mathcal{S}_{\{NID\}}, \mathcal{S}_{\{sd\}}, \mathcal{S}_{\{etld\}}) \times \mathcal{S}_{\{raw\}}$$

The data from the Android phone ($\mathcal{S}_{\{cumulative\ steps\}}$) is a cumulative step number for a given time, so we must first calculate its dynamics (increment for the previous measurement).

$$\begin{aligned} & \mathcal{S}_A(\mathcal{S}_{\{steps\}}, \mathcal{S}_{\{timestamp\}}, \mathcal{S}_{\{date\}}, \mathcal{S}_{\{min\}}, \mathcal{S}_{\{NID\}}, \mathcal{S}_{\{sd\}}, \mathcal{S}_{\{etld\}}, \mathcal{S}_{\{raw\}}) \\ & = \\ & \bigcup_{i=1}^n \left(\mathcal{S}_A^i(\mathcal{S}_{\{cumulative\ steps\}}) - \left\{ \begin{array}{l} 0, \text{ if } \neg \exists \mathcal{L}_A^{i-1} \vee \left(\mathcal{S}_A^{i-1}(\mathcal{L}_{\{date\}}) \neq \mathcal{S}_A^i(\mathcal{S}_{\{date\}}) \right) \vee \left(\mathcal{S}_A^{i-1}(\mathcal{S}_{\{NID\}}) \neq \mathcal{S}_A^i(\mathcal{S}_{\{NID\}}) \right) \\ \mathcal{S}_A^{i-1}(\mathcal{S}_{\{cumulative\ steps\}}), \text{ if } \exists \mathcal{S}_A^{i-1} \wedge \left(\mathcal{S}_A^{i-1}(\mathcal{S}_{\{date\}}) = \mathcal{S}_A^i(\mathcal{S}_{\{date\}}) \right) \wedge \left(\mathcal{S}_A^{i-1}(\mathcal{S}_{\{NID\}}) = \mathcal{S}_A^i(\mathcal{S}_{\{NID\}}) \right) \end{array} \right\} \right) \\ & \quad \times (\mathcal{S}_{\{timestamp\}}, \mathcal{S}_{\{date\}}, \mathcal{S}_{\{min\}}, \mathcal{S}_{\{NID\}}, \mathcal{S}_{\{sd\}}, \mathcal{S}_{\{etld\}}, \mathcal{S}_{\{raw\}}) \end{aligned}$$

Finally, we generate a common large data source from many individual files:

$$\mathcal{S}_A \left(\mathcal{S}_{\{steps\}}, \mathcal{S}_{\{timestamp\}}, \mathcal{S}_{\{date\}}, \mathcal{S}_{\{min\}}, \bigcup_{\mathcal{S}_{\{NID\}}}^{all} \mathcal{S}_{NID}, \mathcal{S}_{\{sd\}}, \mathcal{S}_{\{etld\}}, \mathcal{S}_{\{raw\}} \right) \cup \mathcal{S}_{iOS} \left(\mathcal{S}_{\{steps\}}, \mathcal{S}_{\{timestamp\}}, \mathcal{S}_{\{date\}}, \mathcal{S}_{\{10min\}}, \bigcup_{\mathcal{S}_{\{NID\}}}^{all} \mathcal{S}_{NID}, \mathcal{S}_{\{sd\}}, \mathcal{S}_{\{etld\}}, \mathcal{S}_{\{normalized\}} \right)$$

We generate 10-minute time intervals data from the activity tracker data. Data from the android phone Time property with minute accuracy is 10 minutes raw data, must be normalized as 10-minute accuracy data. The timekey value $\mathcal{S}_{\{tkv\}}^i = Nr(\mathcal{S}_{\{min\}}^i) * 144$ will be a real number on the closed interval [0..144], the correspond time key $\mathcal{S}_{\{TK\}}^i = int(Nr(\mathcal{S}_{\{min\}}^i) * 144)$ is the whole part of the real number.

Each $\mathcal{S}_{\{steps\}}$ value must be broken down into the current 10-minute increments and the previous 10-minute increments. This brings out normalized Android bracelet data.

$$\bigcup_{i=1}^n \mathcal{S}_A^i \left(\mathcal{S}_{\{steps\}} * (\mathcal{S}_{\{tkv\}}^i - \mathcal{S}_{\{TK\}}^i) \right) \times \left(\mathcal{S}_{\{timestamp\}}, \mathcal{S}_{\{date\}}, \mathcal{S}_{\{TK\}}, \mathcal{S}_{\{NID\}}, \mathcal{S}_{\{sd\}}, \mathcal{S}_{\{etld\}}, \mathcal{S}_{\{normalized\}} \right) \cup \bigcup_{i=1}^n \mathcal{S}_A^i \left(\mathcal{S}_{\{steps\}} * (1 - (\mathcal{S}_{\{tkv\}}^i - \mathcal{S}_{\{TK\}}^i)) \right) \times \left(\mathcal{S}_{\{timestamp\}}, \mathcal{S}_{\{date\}}, \mathcal{S}_{\{TK\}} - 1, \mathcal{S}_{\{NID\}}, \mathcal{S}_{\{sd\}}, \mathcal{S}_{\{etld\}}, \mathcal{S}_{\{normalized\}} \right)$$

Data from the iPhone is 10-minute accuracy normalized data ($\mathcal{S}_{\{10minNS\}}$). The time key can be derived with the $\mathcal{S}_{\{tkv\}} = Nr(\mathcal{S}_{\{10minNS\}}) * 144$ calculation.

We used a hash function (CRC32) on the Neptun identifier ($hash(\mathcal{S}_{\{NID\}}) = \mathcal{S}_{\{PK\}}$) before the step data is placed on the intermediate storage server created for our research as an excel table (10minsSteps.xlsx)

The result is the 10-minute normalized step data. $\mathcal{S}(\mathcal{S}_{\{10minNS\}}, \mathcal{S}_{\{DK\}}, \mathcal{S}_{\{TK\}}, \mathcal{S}_{\{PK\}}, \mathcal{S}_{\{sd\}}, \mathcal{S}_{\{etld\}})$

4.4.2 Necessary/Existing Dimensionality Survey

D: dimDate (intermediate storage) unfolded hierarchical date dimension

- id ($\mathbf{D}_{\{id\}}$): unique identifier, continuous serial number {1..∞}
- DateKey ($\mathbf{D}_{\{DK\}}$): unique identifier; serial numbers of year, month, day with leading zeros, in {yyyy}{mm}{dd} composition
- Date ($\mathbf{D}_{\{date\}}$): (the number of days passed since 1900.01.00) in Microsoft date format
- Local Date String ($\mathbf{D}_{\{lds\}}$): year, month, day with leading zeros, in Hungarian date format, in {yyyy}.{mm}{dd} composition
- Year $\mathbf{D}_{\{year\}}$: year identifier {yyyy}
- MonthNr $\mathbf{D}_{\{monthNr\}}$: month serial number {m}
- DayNr $\mathbf{D}_{\{dayNr\}}$: day serial number within month {d}
- MonthStrEn $\mathbf{D}_{\{monthStrEn\}}$: month name in English
- MonthStrHu $\mathbf{D}_{\{monthStrHu\}}$: month name in Hungarian
- MonthStrEnS $\mathbf{D}_{\{monthStrEnS\}}$: month abbreviation in English
- MonthStrHuS $\mathbf{D}_{\{monthStrHuS\}}$: month abbreviation in Hungarian
- Day of Week $\mathbf{D}_{\{DoW\}}$: serial number of weekdays {Monday, Tuesday, ..., Sunday} -> {1..7}
- WeekdayEn $\mathbf{D}_{\{weekdayEn\}}$: days in English
- WeekdayHu $\mathbf{D}_{\{weekdayHu\}}$: days in Hungarian
- DayTypeEn $\mathbf{D}_{\{daytypeEn\}}$: {weekday, weekend}
- DayTypeHu $\mathbf{D}_{\{daytypeHu\}}$: {hétköznep, hétvége}
- Day of Year $\mathbf{D}_{\{DoY\}}$: serial number of day of year {1..366}
- QuarterNr $\mathbf{D}_{\{quarterNr\}}$: serial number of quarters
- QuarterStrEn $\mathbf{D}_{\{quarterStrEn\}}$: quarter in English in Q{quarterNr} composition
- QuarterStrHu $\mathbf{D}_{\{quarterStrHu\}}$: quarter in Hungarian in {quarterNr}. negyedév composition

- WeekNr $D_{\{weekNr\}}$: serial number of weeks {1..52}
- WeekStrEn $D_{\{weekStrEn\}}$: week in English in $W\{weekNr\}$ composition
- WeekStrHu $D_{\{weekStrHu\}}$: week in Hungarian in $\{weekNr\}$. hét composition

T: dimTime (intermediate storage) unfolded hierarchical time dimension

- TimeKey ($T_{\{TK\}}$): {0..143}, the i^{th} 10-minute intervals of the day
- 10mins ($T_{\{10mins\}}$): 10-minute duration in {h}:{mm}-{h}:{mm} composition
- 30mins ($T_{\{30mins\}}$): 30-minute duration in {h}:{mm}-{h}:{mm} composition
- hours ($T_{\{hours\}}$): hourly duration in {h}:{mm}-{h}:{mm} composition

P: dimPerson (intermediate storage) Person dimension

- PersonKey ($P_{\{PK\}}$): students' hashed Neptun identifier
- GenderEn ($P_{\{GenderEn\}}$): students' gender in English {Male, Female}
- GenderHu ($P_{\{GenderHu\}}$): students' gender in Hungarian {Férfi, Nő}

I: dimIntensity (intermediate storage) motion intensity dimension

- StepSumCategoryEn ($P_{\{sscEn\}}$): motion intensity in English
 - {Basal activity, Limited activity, Low activity, Somewhat active, Active, Highly active}
- StepSumCategoryHu ($P_{\{sscHu\}}$): motion intensity in Hungarian
 - {Alapvető aktivitás, Mérsékelt aktivitás, Alacsony aktivitás, Közepes aktivitás, Magas aktivitás, Nagyon magas aktivitás}
- DailyStepSumRange ($P_{\{dssr\}}$): the ranges of the daily step category
 - $0 \leq \text{DailyStepSum} < 2500$
 - $2500 \leq \text{DailyStepSum} < 5000$
 - $5000 \leq \text{DailyStepSum} < 7500$
 - $7500 \leq \text{DailyStepSum} < 10000$
 - $10000 \leq \text{DailyStepSum} < 12500$
 - $12500 \leq \text{DailyStepSum}$
- 10minStepSumRange ($P_{\{10mssr\}}$): the ranges of the 10-minute step category
 - $0 \leq 10\text{minStepSum} < 250$
 - $250 \leq 10\text{minStepSum} < 500$
 - $500 \leq 10\text{minStepSum} < 750$
 - $750 \leq 10\text{minStepSum} < 1000$
 - $1000 \leq 10\text{minStepSum} < 1250$
 - $1250 \leq 10\text{minStepSum}$

4.5 Integration

During integration phase, we describe the production of fact tables and dimensions specified in the ideal data model to be used to answer the questions.

We determine indicators and dimensions needed for integration (not necessarily in this order), but as a result of integration, these should be a kind of documentation.

Finally, we determine the steps of the data loading process (ETL/ELT), looking at their sequence. Our strategies to achieve our integration goal are top-down (ideal model -> source) and bottom-up (source -> ideal model) strategies, both are widely used in information processing and knowledge ordering, in practice, they can be seen as a style of thinking, teaching, or leadership.

4.5.1 Indicators to be calculated to produce the fact tables:

10minNormalizedStepSum: $10minNS^{\{steps\}}(P_{\{PK\}}, D_{\{DK\}}, T_{\{TK\}}, I_{\{10mIK\}})$

DailySteps: $DS^{\{steps\}}(P_{\{PK\}}, D_{\{DK\}}, I_{\{DIK\}})$

Number of students: $St^{\{persons\}}(P_{\{gender\}}, D_{\{DK\}})$

Number of active days: $Activity^{\{days\}}(P_{\{PK\}}, D_{\{MK\}}, I_{\{DIK\}})$

Average daily steps by gender and total: $ADS_{\Sigma\{gender\}}^{\{steps\}}(D_{\{DK\}})$

4.5.2 Additional dimension keys to produce

Daily intensity key: $(I_{\{DIK\}})$

The basis for the categorization is the total number of daily steps of the person under investigation $DS^{\{step\}}(P_{\{PK\}}, D_{\{DK\}}) \geq \{0, 2500, 5000, 7500, 10000, 12500\} \Rightarrow \{0, 1, 2, 3, 4, 5\}$, the necessary and sufficient dimensionality of the indicator is $(P_{\{PK\}}, D_{\{DK\}})$ (Tudor-Locke & Bassett 2004; Tudor-Locke et al. 2011). The result of the logical test is to categorize the number of daily steps of the examined person.

4.5.3 Necessary integration dimensions

T: Time (dimTime)

- TimeKey ($T_{\{TK\}}$): $\{0..143\}$, the i^{th} 10-minute intervals of the day

4.5.4 The data load ETL process

During this process, we load the **S** (Steps) relation properties of the source system and match the dimension keys of the fact table that contains the 10-minute normalized steps in the OLAP system in **Table 8**.

Table 8: 10-minute normalized steps' property mapping.

OLTP system (extract)	transform	OLAP system (load)
$S_{\{10minNS\}}$	\Rightarrow	$10minNS^{\{step\}}$
$S_{\{DK\}}$	\Rightarrow	$D_{\{DK\}}$
$S_{\{TK\}}$	\Rightarrow	$T_{\{TK\}}$
$S_{\{PK\}}$	\Rightarrow	$P_{\{TK\}}$

$$S(S_{\{10minNS\}}, S_{\{date\}}, S_{\{TK\}}, S_{\{NID\}}) \xRightarrow{etl} 10minNS^{\{step\}}(P_{\{PK\}}, D_{\{DK\}}, T_{\{TK\}})$$

After the base ETL we load the dimensions (**Tables 9–11**) defined in the ideal data model and make the necessary conversions.

Table 9: Person dimension's property mapping.

OLTP system (extract)	transform	OLAP system (load)
$P_{\{PK\}}$	\Rightarrow	$P_{\{PK\}}$
$P_{\{GenderEn\}}$	\Rightarrow	$P_{\{gender\}}$

$$dimPerson(P_{\{PK\}}, P_{\{GenderEn\}}) \xRightarrow{etl} dimPerson(P_{\{PK\}}, P_{\{gender\}})$$

Table 10: Date dimension's property mapping.

OLTP system (extract)	transform	OLAP system (load)
$D_{\{DK\}}$	\Rightarrow	$D_{\{DK\}}$
	$left(D_{\{DK\}}, 6)$	$D_{\{MK\}}$
$D_{\{DOW\}}$	$D_{\{DOW\}} \& " - " \& D_{\{weekdayEn\}}$	$D_{\{weekday\}}$
$D_{\{weekdayEn\}}$		

$$dimDate(D_{\{DK\}}, D_{\{DOW\}}, D_{\{weekdayEn\}}) \xRightarrow{etl} dimDate(D_{\{DK\}}, D_{\{weekday\}}, D_{\{MK\}})$$

Table 11: Month dimension-hierarchy's property mapping.

OLTP system (extract)	transform	OLAP system (load)
$D_{\{DK\}}$	$\text{left}(D_{\{DK\}}, 6)$	$DM_{\{MK\}}$
$D_{\{monthStrEn\}}$	\Rightarrow	$DM_{\{month\}}$

$$\text{dimDate}(D_{\{DK\}}, D_{\{monthStrEn\}}) \xRightarrow{etl} \text{dimDateMonth}(DM_{\{MK\}}, DM_{\{month\}})$$

After we have made the basic etl of the DateMonth dimension, the number of rows is related to DateKey granularity with monthly duplicated values, so we must deduplicate the rows noted as below in **Table 12**.

$$\text{dimDateMonth}(DM_{\{MK\}}, DM_{\{month\}}) = \text{dimDateMonth}\left(\begin{matrix} \{DM_{\{MK\}}, DM_{\{month\}}\} \\ \sum \\ \{DM_{\{MK\}}, DM_{\{month\}}\} \end{matrix}\right)$$

Table 12: Walk intensity dimension's property mapping.

OLTP system (extract)	transform	OLAP system (load)
$I_{\{IK\}}$	\Rightarrow	$I_{\{IK\}}$
$I_{\{IK\}}$ $D_{\{sscEn\}}$	$I_{\{IK\}} \&^u - \& D_{\{sscEn\}}$	$I_{\{dsc\}}$

$$\text{dimIntensity}(I_{\{IK\}}, I_{\{sscEn\}}) \xRightarrow{etl} \text{dimIntensity}(I_{\{IK\}}, I_{\{dsc\}})$$

We create fact tables defined in the ideal data model with data manipulation in our data warehouse.

4.5.5 ftDailyActivity: Daily activity fact table $Activity^{\{day\}}(P_{\{PK\}}, (D_{\{MK\}}, I_{\{DIK\}}))$

Daily steps: ($DS^{\{steps\}}$): 10-minute normalized steps must be summarized up through Time dimension to get daily steps and extend the indicator with the daily intensity key.

$$DS^{\{steps\}}(P_{\{PK\}}, D_{\{DK\}}, I_{\{DIK\}}) = 10mNS^{\{step\}}(P_{\{PK\}}, D_{\{DK\}}) \sum_{T_{\{TK\}}}^{all} \times I_{\{DIK\}}$$

Number of active days: $Activity^{\{day\}}$. We must count the daily step related days in the month.

$$Activity^{\{day\}}(P_{\{PK\}}, D_{\{MK\}}, I_{\{DIK\}}) = \text{count } DS^{\{steps\}}\left(P_{\{PK\}}, \sum_{D_{\{DK\}}}^{D_{\{MK\}}}, I_{\{DIK\}}\right)$$

4.5.6 ftDailySteps: Daily steps fact table $ADS_{P_{\Sigma\{gender\}}}^{\{steps\}}(D_{\{DK\}})$

Average daily steps by gender and total: $ADS_{P_{\Sigma\{gender\}}}^{\{steps\}}$

First, we summarize up the 10-minute normalized through Time dimension to get daily steps:

$$DS^{\{steps\}}(P_{\{PK\}}, D_{\{DK\}}) = 10mNS^{\{step\}}\left(P_{\{PK\}}, D_{\{DK\}}, \sum_{T_{\{TK\}}}^{all}\right)$$

4.5.7 Method A (work with one indicator at a time)

Daily steps must be summarized up through Person dimension from PersonKey to gender level.

$$DS^{\{steps\}}(P_{\{gender\}}, D_{\{DK\}}) = DS^{\{steps\}}\left(\sum_{P_{\{PK\}}}^{P_{\{gender\}}}, D_{\{DK\}}\right)$$

Pivoting the daily step indicator with gender attribute.

$$DS_{P_{\{gender\}}}^{\{steps\}}(D_{\{DK\}}) = DS^{\{steps\}}(\odot P_{\{gender\}}, D_{\{DK\}})$$

Calculate the gender independent daily step indicator summary.

$$DS^{\{step\}}(D_{\{DK\}}) = \sum_{P_{\{gender\}}}^{all} DS_{P_{\{gender\}}}^{\{step\}}(D_{\{DK\}})$$

Combine the three daily step indicators through the common DateKey into one fact table.

$$DS_{P_{\Sigma\{gender\}}}^{\{steps\}}(D_{\{DK\}}) = DS_{P_{\{gender\}}}^{\{steps\}}(D_{\{DK\}}) \times DS^{\{step\}}(D_{\{DK\}})$$

Students must be counted up through Person dimension from PersonKey to gender level.

$$St_{P_{\{gender\}}}^{\{pers\}}(P_{\{gender\}}, D_{\{DK\}}) = \text{count } DS^{\{steps\}} \left(\sum_{P_{\{PK\}}}^{P_{\{gender\}}}, D_{\{DK\}} \right)$$

Pivoting the student number indicator with gender attribute.

$$St_{P_{\{gender\}}}^{\{pers\}}(D_{\{DK\}}) = St^{\{pers\}}(\odot P_{\{gender\}}, D_{\{DK\}})$$

Calculate the gender independent student number indicator summary.

$$St^{\{person\}}(D_{\{DK\}}) = \sum_{P_{\{gender\}}}^{all} St_{P_{\{gender\}}}^{\{person\}}(D_{\{DK\}})$$

Combine the three student number indicators through the common DateKey into one fact table.

$$St_{P_{\Sigma\{gender\}}}^{\{pers\}}(D_{\{DK\}}) = St_{P_{\{gender\}}}^{\{pers\}}(D_{\{DK\}}) \times St^{\{person\}}(D_{\{DK\}})$$

Combine the three daily step indicator fact tables and the three student number indicator fact tables through the common DateKey into one fact table.

$$\begin{pmatrix} DS_{P_{\Sigma\{gender\}}}^{\{steps\}} \\ St_{P_{\Sigma\{gender\}}}^{\{pers\}} \end{pmatrix} (D_{\{DK\}}) = DS_{P_{\Sigma\{gender\}}}^{\{steps\}}(D_{\{DK\}}) \times St_{P_{\Sigma\{gender\}}}^{\{pers\}}(D_{\{DK\}})$$

The last step is to divide the three daily step indicators with the related three student number indicators to get the three average daily step indicators.

$$ADS_{P_{\Sigma\{gender\}}}^{\{steps\}}(D_{\{DK\}}) = \left(\frac{DS_{P_{\Sigma\{gender\}}}^{\{steps\}}}{St_{P_{\Sigma\{gender\}}}^{\{pers\}}} \right) (D_{\{DK\}})$$

4.5.8 Method B (work with many indicators at a time)

First, we aggregate the daily step indicator with sum and count aggregate functions through the Person dimension from PersonKey to gender to get the daily step and student number indicators.

$$\begin{pmatrix} DS^{\{steps\}} \\ St^{\{pers\}} \end{pmatrix} (P_{\{gender\}}, D_{\{DK\}}) = \begin{cases} sum \\ count \end{cases} DS^{\{steps\}} \left(\sum_{P_{\{PK\}}}^{P_{\{gender\}}}, D_{\{DK\}} \right)$$

Next, we unpivot the daily step and student number indicators into value (**V**) and a special attribute (**A**) with values of the name of the unpivoted indicators.

$$V^{\{steps, pers\}} \left(P_{\{gender\}}, D_{\{DK\}}, A_{\{DS, St\}} \right) = \left(\begin{array}{c} DS^{\{steps\}} \\ St^{\{pers\}} \end{array} \right) \left(P_{\{gender\}}, D_{\{DK\}} \right)$$

Next, we combine the gender $P_{\{gender\}}$ and our special attribute $A_{\{DS, St\}}$ values into a new $PxA_{\{gender\} \times \{DS, St\}}$ attribute.

$$V^{\{steps, pers\}} \left(PxA_{\{gender\} \times \{DS, St\}}, D_{\{DK\}} \right) = V^{\{steps, pers\}} \left(P_{\{gender\}} \times A_{\{DS, St\}}, D_{\{DK\}} \right)$$

Pivoting our new $PxA_{\{gender\} \times \{DS, St\}}$ attribute values into our special (V) indicator to get four gender dependent daily step and student number indicators.

$$\left(\begin{array}{c} DS_{P_{\{gender\}}}^{\{steps\}} \\ St_{P_{\{gender\}}}^{\{pers\}} \end{array} \right) \left(D_{\{DK\}} \right) = Values^{\{steps, pers\}} \left(\bigcirc PxA_{\{gender\} \times \{DailySteps, Students\}}, D_{\{DK\}} \right)$$

Calculate gender independent daily step and student number indicators by the summary of the gender dependent daily step and student number indicators and combine these two gender independent indicators, with the four gender dependent indicators.

$$\left(\begin{array}{c} DS_{P_{\Sigma\{gender\}}}^{\{steps\}} \\ St_{P_{\Sigma\{gender\}}}^{\{pers\}} \end{array} \right) \left(D_{\{DK\}} \right) = \left(\begin{array}{c} DS_{P_{\{gender\}}}^{\{steps\}} \\ St_{P_{\{gender\}}}^{\{pers\}} \end{array} \right) \left(D_{\{DK\}} \right) \times \left(\begin{array}{c} \sum_{P_{\{gender\}}}^{all} DS_{P_{\{gender\}}}^{\{steps\}} \\ \sum_{P_{\{gender\}}}^{all} St_{P_{\{gender\}}}^{\{pers\}} \end{array} \right) \left(D_{\{DK\}} \right)$$

The last step is to divide the three daily step indicators with the related three student number indicators to get the three average daily step indicators.

$$ADS_{P_{\Sigma\{gender\}}}^{\{step\}} \left(D_{\{DK\}} \right) = \left(\frac{DS_{P_{\Sigma\{gender\}}}^{\{steps\}}}{St_{P_{\Sigma\{gender\}}}^{\{pers\}}} \right) \left(D_{\{DK\}} \right)$$

4.6 The Multidimensional modeling phase

We build the cube(s) with dimensions, dimension hierarchies and measures. In our example we implemented our galaxy schema (**Figure 2**) in Microsoft PowerBI, as the result of our hybrid methodology. We can see it in **Figure 1**. This data cube is the optimal cube to answer the researchers' questions.

Researchers' questions were:

- The students' daily activity by daily steps intensity categories in March (and the preferred visualizing was table).
- Students' average daily activity in March by category and gender (and the preferred visualizing was table).
- Mens', womens' and all average daily steps by the day of the week in March (and the preferred visualizing was radar chart).

On **Figure 3–Figure 5** we can see the print-screens of the dashboards according to the questions. The data cube with dashboards (Takács, 2018) were implemented in Microsoft PowerBI also.

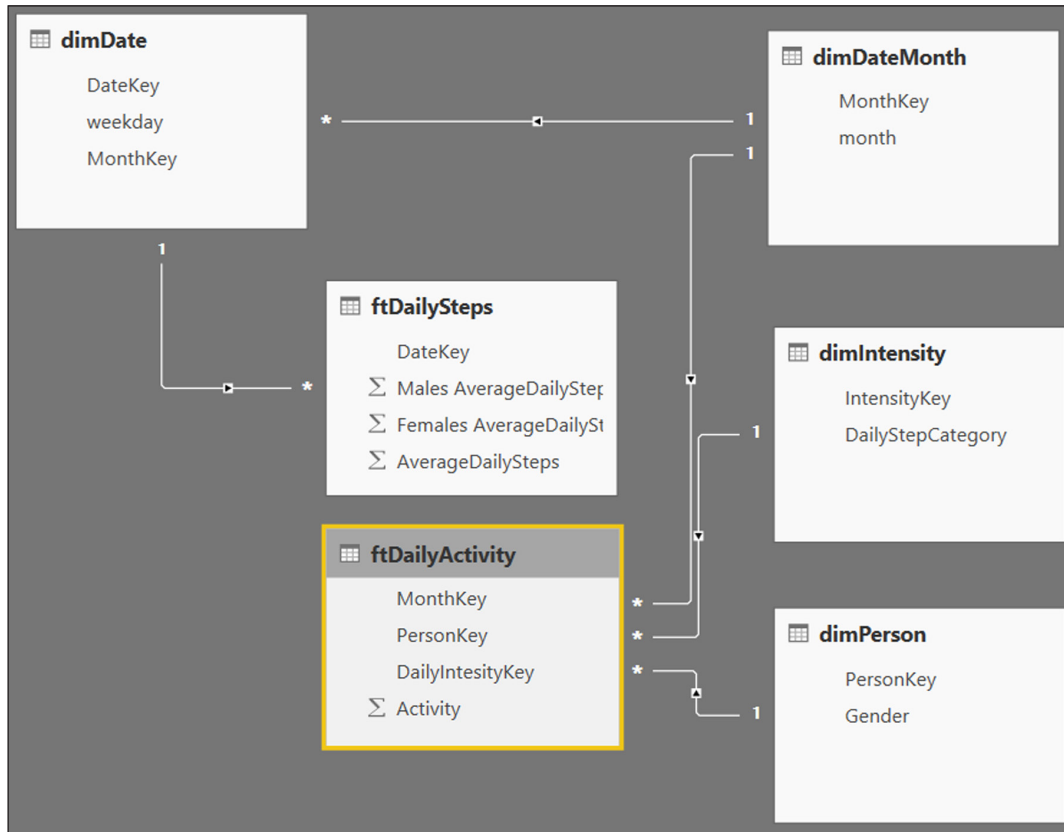


Figure 2: Galaxy schema of the optimal cube.

Student	0 - Basal activity	1 - Limited activity	2 - Low activity	3 - Somewhat active	4 - Active	5 - Highly active	
17662ACA			3	5	5	9	8
2A333AF1		3	6	5	4	7	6
2D1BEEE5			1	4	16	7	3
31CF7C6C			1	2	9	11	8
5760BEB7		1	2	9	7	7	
57718B8		2	3	4	1	2	19
5D804F5E		1	4	7	6	3	10
64429ED1		1		10	2	5	13
66D7C7AA		2	1	3	9	7	9
70FCFA34			1	4	5	9	12
7BF7D822		2	2	11	5	6	4
7F4298ED		3	2	4	3	7	12
8A6C33E3				4	8	6	13
8AC452AF		2	3	9	5	6	5
92E9224D		2	4	5	10	9	1
A1955504		3	2	4	3	7	12
AC5DA34		4	12	11	3	1	
ADFB53C			6	6	7	11	1
B95788B4		6	10	6	5		4
B98FFC7A		5	4	7	5	2	5
COEFC6E		1		1	1	5	23
C3790C53			1	6	9	9	6

Figure 3: Table visualization of question 1.

Gender	0 - Basal activity	1 - Limited activity	2 - Low activity	3 - Somewhat active	4 - Active	5 - Highly active
Female	3,58	3,71	5,69	5,69	5,73	8,36
Male	2,42	3,59	5,83	5,29	5,82	9,64

Figure 4: Table visualization of question 2.

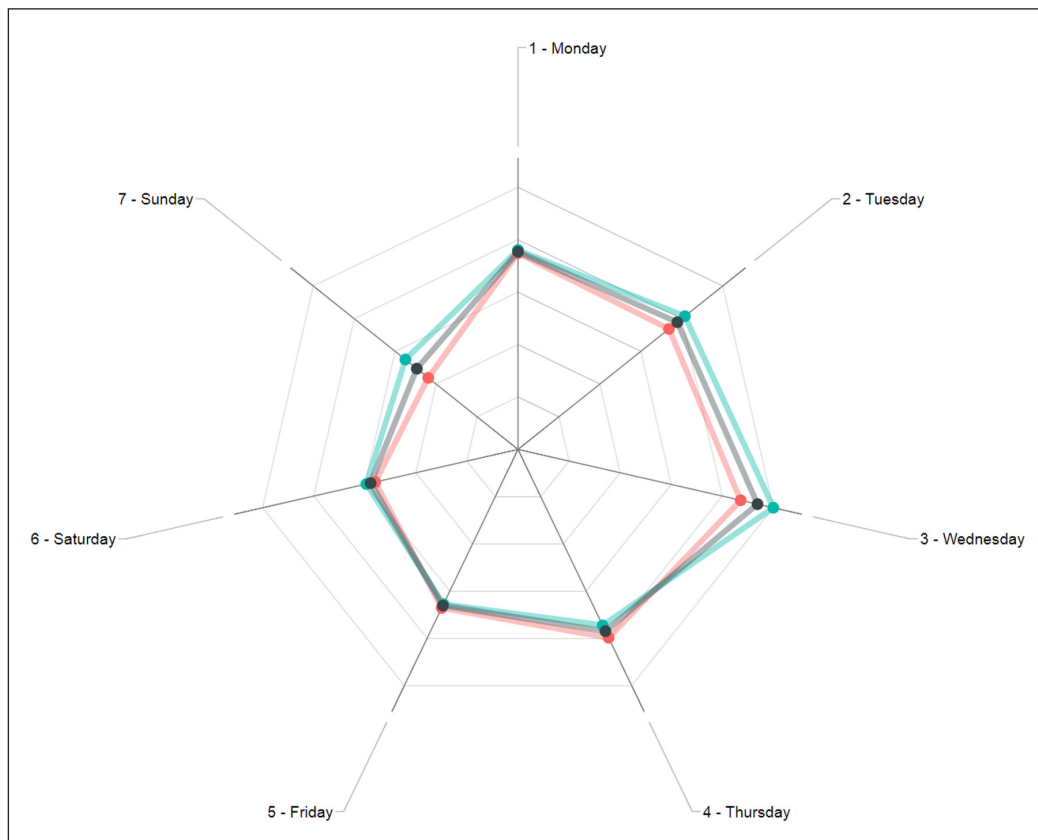


Figure 5: Radar chart visualization of question 3.

5 Summary

In our study we presented a method and concrete designing tool that can decrease a serious deficiency in data warehouse conceptual design phase, when the customer and the vendor should think together to draw up the conceptual plan of a management information system. We provide a kind of ‘business intelligence problem solving thinking’ and a kind of descriptive language that can serve it. We proved with an example, that this approach could work very efficiently in a research area very popular nowadays, that is activity tracking. The problem we presented was simple and there were minimal quantity of management questions, but this hybrid conceptual modeling works in the same way during the conceptual design of a more complex management information system, the visual version of the design process of our example (Takács, 2019) results a very complex graph. The thinking method and the formalisation helps to describe the managerial questions exactly in the conceptual design phase, so it could be an effective intermediate language between designers and creators of the management information system in order to implement successfully, and in the long run help to supply the management or researchers with usual and correct information about their company or research.

Our method has a limitation related to the ETL process, because we focused on the transformation made after the extract-load processes first in the intermediate storage, and last in our Research Data Warehouse. In this example we are not defined notations for complex transformation of ETL process.

Funding Information

The publication is supported by the GINOP-2.3.2-15-2016-00005 project. The project is co-financed by the European Union under the European Regional Development Fund.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

This is a collaborative research in which all authors contributed equally to almost all parts of the manuscript. Viktor László Takács participated in the 2010s data warehouse projects were realized in Hungarian higher education, related to this early management question stenography were developed by Viktor László Takács and Katalin Bubnó in 2012. Gergely Gábor Ráthonyi, Éva Bácsné Bába and Róbert Szilágyi started a research

of physical activity of university students in 2018 and specified the research environment. Viktor László Takács, Katalin Bubnó, Gergely Gábor Ráthonyi and Róbert Szilágyi improved the early stenography to a hybrid modeling method in 2018 work closely to the activity researchers.

References

- Bánné Varga, G.** 2012. *Az adattárház-készítés technológiája*. Budapest: Typotex. (In Hungarian).
- Di Tria, F, Lefons, E and Tangorra, F.** 2017. Cost-benefit analysis of data warehouse design methodologies. *Information Systems*, 63: 47–62. DOI: <https://doi.org/10.1016/j.is.2016.06.006>
- European Parliament and the Council of the European Union.** 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing *Directive 95/46/EC* (General Data Protection Regulation). Bruxelles, Belgium: European Parliament: Council of the European Union.
- Halassy, B.** 1994. *Az adatbázis-tervezés alapjai és titkai*. Budapest: IDG Hungary, (In Hungarian).
- Henriksen, A, Haugen Mikalsen, M, Woldaregay, AZ, Muzny, M, Hartvigsen, G, Hopstock, LA and Grimsgaard, S.** 2018. Using Fitness Trackers and Smartwatches to Measure Physical Activity in Research: Analysis of Consumer Wrist-Worn Wearables. *Journal of Medical Internet Research*, 20(3): e110. DOI: <https://doi.org/10.2196/jmir.9157>
- Kaewkannate, K and Kim, S.** 2016. A comparison of wearable fitness devices. *BMC Public Health*, 16(1): 433. DOI: <https://doi.org/10.1186/s12889-016-3059-0>
- Karapanos, E, Gouveia, R, Hassenzahl, M and Forlizzi, J.** 2016. Wellbeing in the making: peoples' experiences with wearable activity trackers. *Psychology of Well-Being*, 6(1): 4. DOI: <https://doi.org/10.1186/s13612-016-0042-6>
- Kimball, R and Ross, M.** 2013. *The Data Warehouse Toolkit: The Definitive Guide To Dimensional Modeling*. Hoboken, New Jersey: Wiley.
- Krauth, P.** 2008. Üzleti informatika. In: Dömölki, B *Égen-Földön Informatika*. Budapest: Typotex, pp. 549–587. (In Hungarian).
- Lee, I-M, Shiroma, EJ, Lobelo, F, Puska, P, Blair, SN, Katzmarzyk, PT and Lancet Physical Activity Series Working Group.** 2012. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *Lancet*, 380(9838): 219–229. DOI: [https://doi.org/10.1016/S0140-6736\(12\)61031-9](https://doi.org/10.1016/S0140-6736(12)61031-9)
- Macedo, H and Oliveira, J.** 2015. A linear algebra approach to OLAP. *Formal Aspects of Computing*, 27(2): 283–307. DOI: <https://doi.org/10.1007/s00165-014-0316-9>
- Maher, C, Ryan, J, Ambrosi, C and Edney, S.** 2017. Users' experiences of wearable activity trackers: a cross-sectional study. *BMC Public Health*, 17(1): 880. DOI: <https://doi.org/10.1186/s12889-017-4888-1>
- Rizzi, S.** 2008. Conceptual Modeling Solutions for the Data Warehouse. In: Wang, J (ed.), *Data Warehousing And Mining: Concepts, Methodologies, Tools, And Applications*. Hershey, PA: IGI Global. DOI: <https://doi.org/10.4018/978-1-59904-951-9.ch016>
- Takács, V.** 2018. Activity tracking example dashboard. Available at URL [September 2018]. <https://app.powerbi.com/view?r=eyJrljoiZTkxZTEwMmYwLTkNTkEYUWZDFiMTAxOTUwLWliwidiCI6jhmMDcxYjhlLWFjZTMtNGZhNS05MDc3LTAwODRjOTJhMDE5NSIsImMiOj99>
- Takács, V.** 2019. Activity tracking example design process visualization (presliced) at URL [March 2019]. <https://app.powerbi.com/view?r=eyJrljoiMjUyYTBjNjUtMGYyMC00Njg5LWUyYjltNzk0NTU5ZDVkNWU1liwidiCI6jhmMDcxYjhlLWFjZTMtNGZhNS05MDc3LTAwODRjOTJhMDE5NSIsImMiOj99>
- Takács, V and Bubnó, K.** 2012. Felsőoktatási adattárház-tervezés koncepcionális modellje. In: *Technical Reports*, 11, Debrecen: University of Debrecen, Institute of Mathematics and Faculty of Informatics.
- Tudor-Locke, C and Bassett, DR.** 2004. How many steps/day are enough? Preliminary pedometer indices for public health. *Sports Medicine*, 34(1): 1–8. DOI: <https://doi.org/10.2165/00007256-200434010-00001>
- Tudor-Locke, C, Craig, CL, Brown, WJ, Clemes, SA, De Cocker, K, Giles-Corti, B, Hatano, Y, Inoue, S, Matsudo, SM, Mutrie, N, Oppert, J-M, Rowe, DA, Schmidt, MD, Schofield, GM, Spence, JC, Teixeira, PJ, Tully, MA and Blair, SN.** 2011. How many steps/day are enough? For adults. *International Journal of Behavioral Nutrition and Physical Activity*, 8(1): 79–95. DOI: <https://doi.org/10.1186/1479-5868-8-79>
- World Medical Association.** 2013. Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *JAMA: Journal of the American Medical Association*, 310(20): 2191–2194. DOI: <https://doi.org/10.1001/jama.2013.281053>


How to cite this article: Takács, VL, Bubnó, K, Ráthonyi, GG, Bába, EB and Szilágyi, R. 2020. Data Warehouse Hybrid Modeling Methodology. *Data Science Journal*, 19: 38, pp. 1–23. DOI: <https://doi.org/10.5334/dsj-2020-038>

Submitted: 08 October 2018

Accepted: 27 April 2020

Published: 13 October 2020

Copyright: © 2020 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 