CHALLENGES OF SELF-SERVICE BUSINESS INTELLIGENCE

Johannessen, Trond Vegard, NHH Norwegian School of Economics, Helleveien 30, NO-5045 Bergen, Norway, trond.johannessen@nhh.no
Fuglseth, Anna Mette, NHH Norwegian School of Economics, Helleveien 30, NO-5045 Bergen, Norway, anna.mette.fuglseth@nhh.no

Abstract:
The purpose of our research is to enhance the understanding of possible negative effects of business users’ application of self-service business intelligence (BI) tools to support decision making. Self-service BI is an approach to data analytics that is supposed to enable business users to catch, prepare and analyse corporate and environmental data without the involvement of computer specialists. The current literature on self-service BI emphasises the design of new functionalities and user-friendly features. There is little discussion of the qualifications needed to apply the new functionalities purposefully in decision processes. We report on an analysis of 70 bachelor-student groups’ application of a “user-friendly” self-service BI system. We found a significant relationship between the students’ knowledge of data modelling and their ability to apply the system. Inconsistencies in the development of a data cube based on various sources led to serious errors when the data were analysed to support decision making. Implications for the education of users of self-service BI are discussed.

Key words: Business intelligence, self-service business intelligence, data modelling

1 INTRODUCTION

Developments within information and communication technology (ICT) have made it easy to catch data from various sources. Software vendors have developed “user-friendly” systems to support the analysis of data. These developments have enhanced decision makers’ possibilities to base their decisions on data – and thus to question their intuitions and to challenge their biases (McAfee & Brynjolfsson, 2012) – and to improve the effectiveness of their decisions.

Business intelligence (BI) denotes a broad category of applications, technologies, and processes for gathering, storing, accessing, and analysing data to help business users make better decisions (Watson, 2009). The term is usually applied in connection with the use of data that are stored in organisational databases and/or data warehouses. The quality and reliability of such data are under the control of ICT specialists.

Self-service BI (SSBI) is an approach to BI that is supposed to enable business users to search, gather, store, access and analyse data without the involvement of ICT specialists. SSBI is supported by software that allows the users to import data found, for example, on the web. Thus, SSBI makes it possible for decision makers to detect data that reflect sudden and subtle changes in their environments and to analyse such data without waiting for support from ICT specialists.

A literature review on SSBI reveals that current research emphasises the design of sophisticated new functionalities and “user-friendly” features in SSBI. There is little discussion of the requirements needed to apply the new tools in managerial decision making.

Thus, this paper addresses the following research questions:

To which extent can business users with poor knowledge of data modelling apply self-service business intelligence tools effectively for decision support?

Which challenges do business users with limited knowledge of data modelling face when they apply self-service business intelligence tools in their decision processes?
In order to answer these research questions, we have conducted an analysis of how bachelor students have handled an assignment that involved the use of SSBI tools to gather, merge and analyse data from various sources in order to support managerial decision making. An important part of the assignment was the development of a data cube to support the analyses of the need for child-care units.

In our analysis of the students’ handling of the assignment, we emphasised the linkages between the students’ knowledge of data modelling, their development of the data cube and the students’ documentation of the data analyses performed on the cube.

The rest of the paper is organised as follows: In the next section, we give an overview of the literature on SSBI and explain essential concepts. Then we present our conceptual model and our research design, and we explain how we have collected and analysed our data. In the following section, we present our findings. The implications of our findings for teaching and training students are discussed, and further areas of research are suggested.

2 LITERATURE REVIEW

In this section, we first explain the concepts of business intelligence (BI) and self-service BI (SSBI), and we highlight the differences between the concepts. Then we review the literature on SSBI in order to position our research.

2.1 Business intelligence and self-service business intelligence

Business intelligence (BI) has been defined in various ways (see Presthus, 2015, for an overview). Common to the definitions is that the purpose of BI is to support business users, that is, both managers and employees, making “better” (e.g. Watson, 2009) or “better and faster” decisions (Chaudhuri, Dayal & Narasayya, 2011). Some definitions emphasise the technology; other definitions comprise both the technology and the use of the technology to support decision making. We agree with Ask (2013) that a distinction should be made between the technology, the architecture, and the use of the technology.

Building on Simon’s model of the managerial decision-making process with phases of intelligence, design, choice, execution and review (Simon, 1960, pp. 2-3), we consider BI to comprise the search and analysis of data from external and internal sources to support decision making. Simon had borrowed the term intelligence from the military and described his first phase as “searching the environment for conditions calling for decisions”.

Figure 1 is a simplified illustration of the architecture to support BI. The purpose of the figure is to compare and contrast the architecture of BI with the architecture of SSBI, not to focus on the details of BI architectures.

The figure shows a decision maker who searches for and analyses data. The data are found in internal and external sources in a variety of formats. The data have been extracted from their original sources, transformed into a unified format and loaded into the data warehouse. The data warehouse contains historic and aggregated data that have been considered relevant for managing the organisation in general. Such data are termed stationary data (Abelló et al., 2013). A set of BI front-end applications are available for the decision maker to access and analyse the data.

In traditional BI, the architecture is typically set up by the ICT department, and the extraction, transformation and loading (ETL) processes are performed by computer specialists. Thus, the quality and the reliability of stationary data are under the control of computer specialists.
Self-service BI (SSBI) is also defined in various ways. For example, Schlesinger and Rahman (2015) emphasise that end users must understand the semantic layer of the organisational data warehouse so that they become less dependent on ICT specialists concerning their use of the data in the data warehouse. The semantic layer is the business representation of data applying terms that should be familiar to the end users. Abelló et al. (2013, pp. 66-67) argue that the key idea of SSBI is to enable non-expert users to include data in their analyses that are not found in the data warehouse. Alpar and Schulz (2016) distinguish among three levels of SSBI. The lowest level refers to usage of data in reports that have been generated. The second level includes access to data at the lowest disaggregated level in the data warehouse. The third level is similar to the description of SSBI by Abelló et al. (2013). This paper builds on the definition by Abelló et al. (2013). Today’s business environments are characterised by global competition, rapid technological and economic changes and cost pressures. In order to manage such environments, managers must be able to detect data that reflect sudden and subtle changes in their environments and to consider such changes in their decision processes. Figure 2 shows a simplified version of an SSBI architecture (based on Abelló et al., 2013):
Figure 2 illustrates that the SSBI architecture is an extension of the BI architecture with situational data. Löser, Hueske and Markl (2008) define situational data as “data that are needed for the decisional processes but are not part of stationary data”. In line with Simon’s (1960) definition of the intelligence phase of the decision-making process we do not believe that decision makers always know which data that are “needed”.

An important aspect of BI is to search for weak signals in order to evaluate whether the data only indicate a temporary disturbance, or whether they indicate a change that should be acted on. Therefore, we define situational data simply as data that have been searched for outside the stationary data in the data warehouse, for example, on the web. The SSBI architecture is set up and applied by the decision maker, that is, the decision makers are performing the ETL processes themselves.

As illustrated in Figures 1 and 2, we consider the key difference between traditional BI and SSBI to be the flexibility of SSBI architectures to include data that the decision makers may find relevant to explore possible threats and opportunities without being dependent on computer specialists (Abelló et al., 2013). The data may be relevant only to explore specific problems, or they may reflect data needs that have not yet been included in the organisational data warehouse.

Most literature on SSBI is related to the technological development that has made it possible to develop SSBI architectures and analytic tools for non-expert users. Some authors discuss the challenges in developing a consistent data cube, and they propose various ways the technology can support the decision makers when they develop data cubes (e.g. Abelló, et al., 2013; Varga, Romero, Pedersen & Thomsen, 2014). Schlesinger and Rahman (2015) discuss requirements to the analytic tools and the description of the data (metadata) in order to support end users. Abelló et al. (2013, p. 82) admit, however, that many research challenges remain on the way toward SSBI for non-expert users.

Other authors are concerned with the risks related to SSBI and propose various ways to restrict user access to data. For example, Alpar and Schulz (2016, p. 153) propose to divide users into power users and casual users and control the access to data so that only power users can “produce” data. Furthermore, they propose that certain minimum data quality requirements must be decided on (Alpar & Schulz, 2016, p. 154).

In our opinion, control of access to data is not possible when the users are managers at the strategic and tactical levels. Furthermore, one of the advantages of SSBI is exactly to allow managers, who monitor their environments carefully through a variety of sources, to consider data that may reflect weak signals of environmental changes in their decision making.

We believe that SSBI has a potential to improve managerial decision processes. In several studies, we have experienced that managers’ intelligence and analysis processes have increased the effectiveness of their decisions. For example, the CEO of a medium-sized shipping company continuously monitored the markets for weak signals of changes in the demand for transportation services. He prepared the data himself and analysed them using spreadsheets and statistical packages. Due to this monitoring, he was able to act early and reap first-mover advantages on several occasions (Fuglseth, 1989; 2005, p. 293). He also had a comprehensive understanding of the quality of the data that he collected, and he knew which data sources were more reliable than others. This CEO was a business economist with knowledge of data modelling. However, he admitted that the preparation of data for analyses was rather cumbersome with manual entry of some of the variable values. One advantage of the technological development is that more data are available in digital format (McAfee & Brynjolfsson, 2012), and that the possibilities to integrate and analyse data have been greatly extended.

However, referring to a survey by Stodder (2015), Alpar and Schulz (2016, p. 153) maintain that the majority of users are not able to utilise the potential of SSBI to integrate stationary and situational data without the help of BI specialists. As mentioned above, several authors discuss the risks involved in SSBI. The risks are related both to the development of a consistent data cube and to the relevant use of the analytic tools. The discussions of solutions to these problems seem to be more related to restrictions and control than to the education of end users. Alpar and Schulz (2016, p.152) argue that users with “appropriate skills and access rights” should be empowered to move from exploitation of data to exploration of data. However, they do not define or describe what they mean by appropriate skills, and we have not found discussions of requirements to end users in order to apply SSBI architectures and analytic tools appropriately. The purpose of this paper is to inquire into this matter.
3 METHOD

In this section, we first present the conceptual model of our study and our research design. Then we describe how we have collected and analysed our data.

3.1 Conceptual model

Based on the above discussion, we developed a conceptual model as shown in Figure 3.

![Conceptual model](image_url)

Figure 3 Conceptual model

As can be seen in the figure, we assume that the validity of the data analyses performed using SSBI tools is influenced by the consistency of the data cube. By validity of data analyses, we mean that the pivot tables display the perspectives on data as intended by the user. In our study, we had specified certain analyses to be performed by the subjects. Thus, we could control whether the pivot tables were valid displays. A consistent data cube implies that all necessary relationships have been created correctly, requiring that primary and foreign keys are established. Furthermore, we assume that the users’ knowledge of data modelling influences their ability to create a consistent data cube.

3.2 The research design

In order attain insights into possible relationships between knowledge of data modelling and the adequate use of SSBI tools to develop a data cube and analyse data we developed an assignment (a home exam) in a mandatory bachelor course in business data processing. The students were asked to assist the manager of a Norwegian child-care provider that wanted to assess the market opportunities for establishing new child-care units.

The students were second-year students, and 401 students attended the course. The answers to the home exam must be submitted in groups of maximum four students, and the students had 8 ½ days to prepare an answer paper of maximum 30 pages.

As business users of tomorrow, business students are often selected as surrogates for managerial decision makers (Remus, 1986). Results from empirical research suggest caution in using students as surrogates, particularly in tasks that require experience (Hughes & Gibson, 1991). However, in relatively structured contexts and in contexts where age and experience are supposed to have only a moderate effect, the use of students is not considered a serious limitation (Remus, 1986; Weick, 1967; Mortensen, Fisher & Wines, 2012). In our study, the emphasis is on knowledge of data modelling principles, the task is rather structured, and the students are not supposed to make decisions that require experience as a child-care provider. Therefore, we consider the business students to be suitable surrogates for business users.

The assignment consisted of three tasks. The first task required the subjects to demonstrate their understanding of data modelling, that is, to develop an entity-relationship (ER) diagram (Chen, 1976) and relational database tables (Codd, 1972) that satisfied certain requirements to a database for the child-care provider. In the second task, the subjects had to develop a data cube from five data sources. In the third task, the subjects must use the data cube for specific analyses and come up with recommendations for further action. Thus, our research design is a correlational design (Kidder & Judd with Smith, 1986; Shadish, Cook & Campbell, 2002).

Four of the five data sources were available to the students as part of the assignment. These sources were:

- a table with data on the number of children in each type of child-care units for the period 2002 – 2013 (private, municipal, regional and national),
- a table with data on municipalities as per year 2013,
- a text file with data on the population of Norway in each municipality for the period 1986 – 2014,
- a text file with data on the number of child-care units of each type for the period 2002 – 2013.
The fifth source was an html document with data on Norwegian counties that the students had to collect from Wikipedia.

The SSBI tools applied in the course were MS Power Query and MS Power Pivot. These tools are now integrated in MS Excel 2016, and will probably become common end-user SSBI tools (Presthus, 2014). In the course, Power Query was mainly used as an ETL tool to extract, transform and load data into Power Pivot. Power Pivot was then used to build the dimensional data model (the data cube) to analyse data.

Prior to receiving the assignment, the students had two class-room lectures on data modelling and two class-room lectures introducing them to Power Query and Power Pivot. The lectures comprised basic principles of data modelling and an introduction to the development of star schemas with fact and dimension tables (Codd, 1972; Kimball & Ross, 2013). By basic principles of data modelling we mean to analyse a task, develop an entity-relationship (ER) diagram and transfer the diagram into table definitions on third normal form with specifications of primary and foreign keys and other attributes needed to handle the task. Following the lectures, the students had two computer labs where they got practical training in creating data models and two computer labs using the Power BI tools on relevant cases.

The tasks on developing and analysing a data cube (tasks 2 and 3) had been designed so that they required the students to:

1. analyse the tasks and decide which data they would need and along which dimensions they would need to analyse the data,
2. evaluate the contents of each data source and decide how to make the data series comparable, for example, data that had been collected on January 1st, 2010 must be compared with data collected on December 31st, 2009, and not with data collected on December 31st, 2010,
3. use Power Query to
   - extract relevant data from the data sources (not all data were relevant for the task),
   - identify and/or create unique primary and foreign keys according to the analysis of the tasks (Power Pivot does not accept composite keys, i.e. keys consisting of two or more attributes),
   - transform the data into a suitable format, for example, convert data from text strings to numerical values,
   - combine some attributes from two sources before loading the data into Power Pivot,
4. present the data cube in Excel for further analyses of data as illustrated in Figure 4,
5. perform and document specific analyses on the data cube using the pivot-table tools.

Figure 4  Example of presentation of data cube in Excel
3.3 Data collection and analysis

Our collection of data was influenced by the fact that the assignment had been developed as a home exam in a mandatory course with 401 students, and that not all examiners had used the same system for grading the tasks.

A total of 120 exam papers were handed in. The examiners of 70 papers had used the grades A – F for each task, so that 70 papers were suitable for our study. Of these papers, we had access to 20 full answer papers, that is, to the students’ detailed solutions and their argumentations.

We started with the analysis of all 70 papers. We performed a Spearman correlation analysis of the grades obtained for the task on data modelling (task 1) and the grades obtained for the task on the development of the data cube (task 2). We used the grade in task 1 as a proxy of the students’ knowledge of data modelling. The data cube developed in task 2 was graded according to correct application of relevant data modelling principles (Codd, 1972; Chen, 1976; Kimball & Ross, 2013).

We continued with a detailed analysis of the 20 full answer papers:

In these analyses, we used the students with a grade of D and below in task 1 to represent business users with poor knowledge of data modelling (research question 1). We analysed whether and possibly how these students’ lack of understanding of data modelling principles influenced their development of the data cube (task 2) and their analyses (task 3).

We then went through the remaining answers in order to identify the challenges the students had developing the data cube. We analysed the developed data cubes in order to locate errors. The errors in each answer paper were checked against the students’ answer to task 1 in order to enhance our understanding of whether lack of understanding of certain data modelling principles had influenced the development of the data cube (task 2).

Finally, the students’ data analyses in task 3 were compared with the correct solutions to the analyses specified in the assignment. We related the students’ data analyses to the consistency of the data cube and studied how inconsistencies in the data cube influenced the results of their analyses.

4 FINDINGS

We found a significant correlation between the grades of task 1 and task 2, representing knowledge of data modelling and consistency of the developed data cube (R=0.380, p=0.001). Thus, the assumption in our conceptual model (Figure 3) of a relationship between users’ knowledge of data modelling and their ability to create a consistent data cube was supported.

4.1 Addressing research question 1

Of the 20 answers that we analysed in detail, we used three answer papers to address research question 1. These three answers had obtained a grade of D+ or below for the task on data modelling. These student groups had also obtained a grade of D or below for their answers to the task on development of the data cube. The three groups had managed to develop a data cube, and the presentation of the data was similar to the presentation in Figure 4, quite apart from the fact that the data values presented in the pivot tables and diagrams were not correct.

Figure 5 shows a data model for one of the low-scored answers to task 2. The figure illustrates that the students have identified only two dimensions: Time (year) and Geography (municipality, county). Type of child-care unit has not been recognised as a dimension. Furthermore, the table FacBefolkning2002-2013 has one attribute for males and one attribute for females, but the attributes for AntKommunale, AntFylkeskommunale-statlige and AntPrivate have only values for children of both genders. Thus, in this case the number of boys and girls should have been summarised. The result of such weaknesses is that the use of the SSBI tools for relevant analyses becomes rather complex, and the risk of errors increases.
4.2 Addressing research question 2

The remaining 17 answer papers were used to address research question 2. With regard to these 17 papers, the majority of the students had attained relevant knowledge on data modelling to solve task 1. Most students were able to develop a relevant data model with the correct entities and relationships among the entities.

The problems handling task 1 were primarily related to a lack of compliance between the entity-relationship (ER) diagram and the table definitions, for example that a relationship in the ER diagram was not represented by a relationship between a primary key and a foreign key in the corresponding table definitions. Repeated errors in the specification of a consistent set of primary and foreign keys in the table definitions indicate that the students have not quite understood basic principles of data modelling.

11 of the 70 student groups attained the grade A in task 2, that is, about 16%. However, among the 20 full answer papers that we had access to only one group had got an A, see Figure 6.

In the following, we will apply the list presented in section 3.2 to discuss challenges that the 16 student groups who attained the grades B and C had answering tasks 2 and 3:

1. Few of these students had actually managed to analyse the assignment so that they presented the data in dimension and fact tables even though principles and examples had been taught in class. However, most of the students had managed to define primary and foreign keys among the tables so that the data cube gave consistent analyses.

2. Other challenges seemed to be related to the interpretation of entities in order to make data from different data sources comparable. For example, the data source with the number of children in child care had been collected on December 31st each year, whereas the data source containing data about the general population had been collected on January 1st each year. Students who had not detected these differences in the date for collecting data or interpreted the differences in a reasonable way, compared data for the subsequent years instead of the same year. The result was that their analysis provided an incorrect representation of the coverage of child-care units, that is, the number of children in child-care units/the number of children.

3. Most students were able to use Power Query to extract the relevant data from the data sources to handle the analyses required.

Figure 5 Example of dimensional model from a low-scored answer paper
However, the answers revealed a few technical difficulties preparing the data before loading them into Power Pivot, such as:

- splitting a text string and selecting part of the string,
- converting text strings into numerical values,
- identifying and/or creating primary and foreign keys (Power Pivot does not accept composite keys).

Such operations are essential both to establish columns with values that can be used for calculations and aggregations – and to create unique primary and foreign keys that can be used when analysing data in Power Pivot:

When extracting data from Wikipedia to Power Query, the column for population data per county was imported as a string, for example, ‘3&505&287198&287198’. The characters representing the population in the example string above is 287198, which is found between the second and the third ampersand (the & symbol). The selected string then had to be converted into a numerical value.

Some students were not able to create unique primary and foreign keys so that data tables could be joined. For example, data about municipalities and counties were located in two different data sources, and thus had to be joined into a geography dimension. This operation required that the students first defined a common attribute in the two tables. The consequence of not joining the two tables was that the students had problems aggregating data per county in their analyses.

4. As regards the presentation of the data cube in Excel, most student groups had managed to create a presentation similar to the example in Figure 4 (cf. comments to point 1). However, the technical challenges discussed above made it difficult to exploit the full potential of the SSBI tools to analyse data.

5. The most serious challenges were related to column and row operations in Power Pivot. There were few errors in such operations, but when the errors occurred, they had severe consequences for the analyses. Furthermore, the students did not seem to detect such errors when analysing data in Power Pivot. The students commented on results that did not seem logical, such as more children in child-care units than children in the municipality, but instead of looking for errors in
their development of the data cube or their use of the analytic tools, they tried to find arguments supporting their results. Examples of such arguments were that some child-care units probably included children from other municipalities. Seemingly, the students had not checked the operations on a subset of the data, for example, by comparing results from Power Pivot with results obtained using well-known operations in Excel.

5 DISCUSSION AND CONCLUSION

In our study, bachelor students have used SSBI tools such as Power Query and Power Pivot to handle the development of a data cube and analyse data according to certain specifications. Our study has shown that the students did not have problems applying Power Query when extracting relevant data for handling the assignment. A few students had some problems preparing the data before loading them into Power Pivot. These problems were related to the handling of text strings, such as splitting a text string and selecting part of the string, and converting text strings to numerical values. Most students managed to use Power Pivot from a technical perspective.

Our study has demonstrated that end users need basic knowledge of data modelling to be able to develop a consistent data cube that can provide valid analyses. The students that had not attained basic knowledge of how to define primary and foreign keys to combine tables did not manage to develop a consistent data cube, and their data analyses were not correct. In this respect, our study supports the arguments by Stodder (2015) and Alpar and Schulz (2016) regarding the risks related to SSBI. However, instead of emphasising restrictions and control of users, we believe that we should educate end users.

Our study showed that even though few students managed to specify dimensions and fact tables adequately, the majority of the students had gained knowledge of data modelling so that they were able to prepare the data with adequate primary and secondary keys before loading the data into Power Pivot. These students were also able to provide valid analyses of the situation regarding child-care units and coverage of children in the child-care units as specified in the assignment.

A few students with errors in the aggregations of data and in the computations on data did not seem to detect the errors. Some students commented on results that did not seem logical, but they tried to find arguments in support of the results instead of performing manual checks on the validity of their analyses.

Our study has implications for teaching. Even though SSBI tools such as Power Query and Power Pivot are easy to use from a technical perspective, they do not support the users developing a consistent data cube. Without knowledge of how to model a data cube, the use of the powerful operations on a data cube may provide wrong analyses and result in ineffective decisions. Thus, the development of “user-friendly” SSBI architectures and tools does not replace teaching principles of data modelling and training the students in applying the principles on challenging tasks.

In addition to teaching the students principles of data modelling such as table definitions with needed attributes and adequate specifications of primary and foreign keys, the teaching should emphasise and illustrate how the students should analyse a task and develop models of the data cube before starting to develop the cube. The teaching should also include practical advice to the students on how to check the validity of their data analyses.

Furthermore, the training should include tasks designed to challenge the students on logical aspects, such as detecting when data must be adapted in order to be comparable, in addition to technical challenges related to how to present data tables in a data cube so that they can be combined and aggregated as intended. Each training section should be followed up with a thorough debriefing with a discussion of the students’ experiences handling the tasks in order to enhance learning (Savery, 2015).

Our study certainly has limitations. Our findings are based on detailed analyses of 20 answer papers only, and the analyses are based on the outcomes of the students’ efforts. We have not been able to study their processes. Furthermore, the study does not include search for data. All data sources needed to handle the assignment were included as part of the assignment, and all data had been properly described in the assignment. Furthermore, the assignment did not include challenges related to semi-structured or unstructured data. Future studies should be designed to catch users’ decision-making processes or part of
such processes. Examples of such studies could be the study of how users search for data on the web, how users plan to develop and actually develop a data cube – and how they apply the SSBI tools for analyses of the data cube in order to improve decision making.

6 REFERENCES


Exploring and Extending the Legacy of Howard S. Barrows (pp. 5-18). West Lafayette, Indiana: Purdue University Press.


