An Instructional Design for Data Warehousing: Using Design Science Research and Project-based Learning

Roelien Goede
North-West University, South Africa

Abstract

The business intelligence industry is supported by data warehouses. A data warehouse is a database that supports fast queries for strategic decision making. Students should be able to leave university and use their skills to the benefit of an organisation. The aim of design science is to create new knowledge through development of an artifact. The paper reports on a design science research project to create an instructional design for a module in data warehousing. The design of the instructional design is informed by project-based learning. The aim of the module is to equip students for the demands of the business intelligence industry.

1. Introduction and problem statement

Kimball et al. defines business intelligence (BI) as “a generic term to describe leveraging the organization’s internal and external information assets to support improved business decision making” [1]. This definition highlights the reliance of decision making on information or data in an organisation. The data component in a BI system is a data warehouse (DW) [1]. One can argue that DWs drive strategic decision making in large organizations.

Students in the 4th and final year of a Bachelor’s program in Information Technology at a university in South Africa complete two semester modules in DW along with, amongst others, modules in project management and decision support systems to prepare themselves for the BI/DW industry. This preparation should not only be in terms of a sound theoretical base, but it should include application of DW principles to industry type projects. The focus on industry type projects motivates project-based learning (PBL) as a teaching strategy for this module.

The paper reports on the development of an instructional design for the DW modules to achieve theoretical knowledge and practical skills to empower graduates to take up their role in the BI/DW industry.

The instructional design is viewed as an artifact in terms of design science research (DSR). This research process is guided by the phases of DSR.

The paper commences with a discussion of DSR as research methodology in Section 2. DSR uses descriptive and prescriptive knowledge. A short discussion is provided on data warehousing in Section 3 and project-based learning (PBL) in Section 4 as these disciplines provide the descriptive and prescriptive knowledge for this DSR project. Section 5 focusses on the phases of the DSR project and forms the empirical report of the study. Concluding remarks and future work are discussed in Section 6.

2. Design science research methodology

According to Gregor and Hevner, design science research is used for the design and development of socio-technical artifacts such as decision support systems, modeling tools, and evaluation methods for information systems [2]. The methodology is often used in the Engineering discipline where the creation of artifacts is a primary activity. Epistemologically, new knowledge is created through the design and development of the artifact.

In arguing the scholarly value of DSR projects, Gregor and Hevner explains the role of knowledge in DSR [2]. They distinguish between descriptive (Omega) knowledge and prescriptive (Lambda) knowledge.

Descriptive knowledge or “what-knowledge” provides the theoretical context of the artifact. Phenomena and the interactions between phenomena constitute descriptive knowledge.

Prescriptive knowledge describes the “how-knowledge” and provides methodological guidance in the development of the artifact. Figure 1 shows the role of knowledge in a DSR project. The development of the artifact is influenced by the specific application environment (in this instance the specific programme at a specific university) and the human capabilities of the designer (in this instance the lecturer who is also the researcher and author of this paper.)
In this study the instructional design of a module in data warehousing is viewed as an artifact. Descriptive knowledge includes aspects such as:
1. What is a data warehouse?
2. What is project-based learning?
3. What constitutes an instructional design?
Prescriptive knowledge for study includes aspects such as:
1. How should a data warehouse be developed?
2. How should PBL concepts be included in an instructional design?

Peffers et al. [3] presents the following phases in a DSR project.
1. Identify problem and motivate
2. Define objectives of the solution
3. Design and development
4. Demonstration
5. Evaluation
6. Communication

They (Peffers et al.) argue that there are more than one entry point in the DSR process, in other words a DSR project does not need to start at phase one in the list [3].

The role of knowledge in DSR necessitates the discussions of DW and PBL to follow.

3. Data warehousing

This section on data warehousing is divided into typical descriptive and typical prescriptive knowledge of DWs.

3.1. Descriptive data warehouse knowledge

The popular definition of a data warehouse is given by the “father” of data warehousing, W. Inman. He defines a data warehouse as “a subject oriented integrated, non-volatile, and time variant collection of data in support of management decisions” [4]. In order to achieve an understanding of data warehousing a distinction needs to be made between operation information systems and DWs. Operational information systems are the operational systems we find in our everyday life of buying items in supermarkets and using bank teller machines – their aim is to gather data about transactions. They store data in normalized database systems. These systems are for on-line transactional processing (OLTP)

Figure 2 depicts the high level data warehouse architecture [1]. The data warehouse may be viewed in three parts: the back room, presentation server and the front room.

The back room is the technical engine room of the data warehouse Data warehouses use OLTP systems as input, and often more than one OLTP. In DW terms these systems are called source systems. Data from different source systems needs to be integrated (extracted, transformed and loaded or ETL) into a database (presentation server) that is designed to provide fast results to queries. The ETL process can be very complicated as different source systems might have vastly different technical infrastructure such as database engines. Data format also creates integration problems. Kimball et al.
argues that ETL is a substantial part of the overall DW development effort [1].

The presentation server is the main data store of the data warehouse. Data, resulting from the ETL process, is stored in relational tables. The tables are not normalized and organized in a dimensional design or star schema. Dimensional modeling differs substantially from typical entity relational models where tables are normalized. A discussion of dimensional modeling does not fall within the scope of the paper and can be found in [1].

Figure 2. High level data warehouse architecture [1]

Figure 3. Data warehouse design methodology [1]
Dimension tables are designed in a conformed way that it may be used by several star schemas representing different business processes.

The front room services concern the access to the data warehouse data by the business user. Business users are not always aware of the possible uses of data warehouse data [1] [4]. It is the role of the data warehouse team to educate the business users to be able to use the data warehouse to its full potential.

3.2. Prescriptive data warehouse knowledge

There are two well-known methodologies for developing DWs. The first is the data-driven methodology of Inmon described in [4]. Traditional OLTP systems are typically developed by first collecting user requirements of a new system. These requirements are then implemented and tested before a new system is released. Inmon argues that users of DW are not able to provide exact requirement as they are not able to foresee all the opportunities a DW will provide them [4]. Inmon’s methodology focusses on gathering all available data in an organisation before integrating it into a large data store. From this data store business analysts create data marts to be used by different business process managers in an organisation.

Kimball et al. differs from Inmon in that they start with the collection of user requirement [1]. The process suggested by Kimball et al. is depicted in Figure 3 [1]. Business users are interviewed to discover key performance areas where availability of integrated data may benefit the organisation most.

After business requirements are gathered, the methodology involves three parallel tracks. The first is the technical track and involves the technical architecture design and the selection of products.

The second track focuses on dimensional modeling and data staging or better known as the ETL process of extracting (E), transforming (T), and loading (L) the data from the source systems into the presentation server.

The third track focuses on the development of user applications to extract information to support the business needs from the data warehouse.

After the first iteration is implemented a second iteration is started to extend the data warehouse to for the use of other business units. The entire project is managed according to sound project management principles as indicated on Figure 3.

A technical difference between these methodologies that is important for this study is the storage of the integrated data. In Inmon’s methodology data is stored in a central database which is similar to OLTP systems’ databases. In the Kimball et al. methodology data is stored in a dimensional model. There is a substantial difference in the design of a dimensional model compared to an OLTP database.

4. Problem-based learning

The aim of this section is to provide a brief introduction to problem-based learning (PBL). Once again it will be divided into descriptive and prescriptive knowledge. Thomas provides a detailed discussion on PBL and a review of PBL literature [5].

4.1. Descriptive PBL knowledge

PBL is a teaching approach that uses projects to guide learners to discover knowledge about a phenomenon. Helle et al. discusses the following characteristics of PBL [6]:

1. A problem situation drives the learning activity, it has a project orientation.
2. Project-based learning is distinguished from problem-based learning as it is aimed at constructing an artifact.
3. The learner is in control of the learning process.
4. Learning is contextualized in a real problem situation.
5. Students are tasked with creating and using multiple forms of representation of knowledge.
6. Students have improved intrinsic motivation to learn.

PBL may be viewed as a teaching strategy to enhance the self-directed learning skills of students, which in turn lead to the development of students as life-long learners [7].

4.2. Prescriptive knowledge of PBL

It falls outside the scope of the paper to provide a full description of guidelines to use PBL in instructional design. However the guidelines for successful PBL projects provided by Hickey are relevant for this study. A successful PBL project [8]:

1. Completes over a longer period of time (unit, month, quarter or semester);
2. Addresses concepts across the curriculum;
3. Increases rigor and relevance of instruction.
4. Students can arrive at the answers in a variety of methods.
5. Students will come up with a variety of solutions for the project.
6. Students will develop a long-term understanding of the information.
5. An instructional design to prepare data warehouse students for industry

This section forms the empirical report of the study and is organized according to the phases of a DSR project discussed in section 2. The discussion starts with a short description of the context of the study.

5.1. Context of the students

In South Africa a Bachelor’s degree is completed over three years. Before students can enter a Master’s degree, they need to obtain an Honors degree, which is typically a one year full-time study. We regard this as the final year of a four year degree programme. A Bachelor’s degree in Information Technology is a prerequisite for the Honors degree in Information Technology. The Honors degree is presented full-time and part-time with combined contact sessions after office-hours. About a third of the students in the program are working full-time in the information technology industry while they complete their Honors degree part-time over a period of two years.

Two thirds of the students have no working experience and hope to use their Honors degree to enter the industry at a higher level than what they would with their 3 year Bachelor’s degree.

Students who select the module in data warehousing also do a module in advance database systems. They have to select two modules from computer security, project management, decision support systems, and artificial intelligence. They also have to complete a substantial research project to satisfy the requirement of the degree. All the modules are divided into two semester units with a study credit weight of 120 hours per semester and 240 hours per year. The modules all have a single one hour contact session per week in the evening. And the research project requires 360 hours during the year.

5.2. Identify problem and motivate

The problem to be addressed by the artifact is the theoretical nature of data warehousing studies in the controlled environment of the university. Due to a lack of access to industry the module was previously presented mainly from a theoretical perspective with a series of small practical assignments. Such an approach does not provide the student with confidence that he or she is ready to start working in a BI/DW environment after completion of the module. Furthermore part-time students come to the campus after work, many traveling more than 100 km expecting more than a lecturer reciting the contents of a text book.

5.3. Define objectives of the solution

After interpretive interviews with industry professionals it was decided that the educational design should provide students with theoretical knowledge and practical skills.

Theoretical aspects identified coincided greatly with the contents of well-respected industry used publications in the field such as the book by Kimball et al. [1]. Some additional aspects mainly focusing on the work of Inmon was identified [4].

On a practical level, students should be able to implement a small data warehouse from data provided in terms of ETL processes, data modeling, and user access to data.

5.4. Design and development

The design and development of the instructional design was done according to the ideas of PBL and SDL.

The 4th year level of the student meant that they have reached a high level of SDL and since there is only one hour of contact time available per week it was decided to develop a strategy of prepare, attend, and reflect for the theoretical part of the module.

The theoretical component was divided into study units comprising of about 6-8 study hours per week covering the aspects identified in the discussions with industry professionals. Formative evaluations are scheduled and there is a summative examination after each semester. Students are also required to reflect on their practical experience when answering questions on the theoretical aspects.

Formal evaluation prioritizes the skills of the students to develop dimensional models from case studies.

An industry professional provides data and requirements for the practical project. He also assists with the evaluation of the efforts of the students. It was decided to use a hybrid DW design methodology, combining the data-driven and requirements driven ideas from Inmon and Kimball et al. respectively. This is achieved by first giving desensitized industry data to the students for a week before allowing them to interview the industry professional acting as the business user of the data.

The practical project’s requirements were carefully selected to cover as many of the theoretical aspects of the module as possible.

The project has deliverables in terms of software and documentation that must be submitted for evaluation.

The final mark for the module is an equally weighted combination of the marks achieved for the theoretical part of the module and the practical part of the module.
5.5. Demonstration

In terms of the prepare, attend, and reflect strategy adopted for the theoretical part of the module, students are required to create a mind map and answer a set of questions on a study unit prior to attending the contact session.

During the contact session the lecturer acts as facilitator and student questions guide the discussion. Students are encouraged to take part in discussion which is enhanced by input from part-time students working in industry.

After the contact session students have to submit a one page summary reflecting on the contact session, highlighting aspects of DW they learnt.

In terms of the practical component of the module, students divide themselves in teams of three students. The data is large in quantity – typical more than a million records are provided by the industry professional. The data warehouse project they have to complete is divided into milestones.

The first part of the project focuses on the backroom of the data warehouse architecture presented in Figure 2.

1. Convert the received data into a manageable format. Students need to create tables in a database management system before using a bulk loader to load the data. These tables form the source tables for their warehouse.
2. Interview the “business” user and compile user requirements. After creating the tables in the first milestone, the students have a basic understanding of the data and they are able to ask the business user sensible questions. They need to develop a requirements specification document to demonstrate their understanding of the business requirements.
3. After understanding the business requirements, a project plan should be developed. The plan should be according to the methodology of Kimball et al. [1]. Students are required to reflect on their progress on a weekly basis.
4. Create a dimensional model for the DW. Students also have to develop documentation in support of their model. This includes fact table sheets, dimensional model designs, including hierarchies and source to target maps.
5. Create and implement an ETL process to load the data. Students are given data from two source systems that need to be merged. Care should be taken to remove duplicate data and to link data that belong together such of sales of the same product.

The second part of the project focuses on the front room of the data warehouse architecture presented in Figure 2.

6. Create a data warehouse browser in a graphical programming language such as C#. The browser should allow its user to create and execute DW queries. Students have freedom to extend this browser to have more functionality in terms of query generation.
7. Create an online analytical processing (OLAP) cube in the DW tool provided. Operations on OLAP cubes require a clear understanding of hierarchies in the dimensions of the data. Hierarchies are natural groupings in the data such as days that forms a month, months combining into a quarter and quarters forming a year.
8. Link from typical spreadsheet software to the cube and create a dashboard user application. Students are expected to present a creative design demonstrating their understanding of the business requirements.
10. Choose any analytical tool to demonstrate use of the DW for other purposes than the given requirements.

Evaluation of each milestone is done using checklists and providing written comments to each group. The first 5 milestones are completed in the first semester and the last 5 in the second semester.

After the first semester all students are provided with high quality data tables to provide them with a fair opportunity to complete the final milestones.

5.6. Evaluation of the instructional design

This instructional design has been implemented for the past three years at two campuses of the University. Each year about 25 students enroll for the module on each campus resulting in about 150 students completing the module. At the end of every year we give the students a standard university module evaluation to complete as well as an open-ended set of questions.

The students enjoy the module and mostly give positive feedback. Complaints are usually connected to lack of assistance to solve specific problems during the practical project.

Students also report that the workload of this module is higher than their other modules. They have problems with their time management and complain that they run out of time at the end of the semester.

It is indeed difficult for the students to create a project plan before finishing the theoretical part of the module. It is difficult for them to estimate the
effort of the activities since they do not know beforehand what each activity involves. In order to guide them, we expect them to give more presentations during the semester to demonstrate and compare their progress with that of the other students in the class.

There is a procedure for the students to ask the industry-professional technical questions. Very few students make use of this procedure.

More than 50% of all our Honors in Information Systems students find their first employment position in the BI/DW industry. The University is often approached by industry aiming to recruit our DW students.

5.7. Communication

Communication in terms of DSR focusses on scholarly communication as a means of distributing new knowledge.

This paper forms part of this communication. The study guide for the module in question is also a form a communication within the university community. Faculty members and external moderators are invited to attend the final demonstration of the student projects.

This instructional design is internally available in to the faculty members in the form of the study guide for the module. Other modules may be redesigned according to similar PBL principles by using this study guide as an example.

5. Conclusion and future work

Project-based learning (PBL) provides a teaching strategy to bring industry into the classroom. The instructional design reported in this paper is especially successful in this regard due to the commitment of an industry professional. His presents not only results in realistic data and requirements but also provide an employment opportunity for some of the students. The industry professional views his contact with the students as a year-long interview.

The study demonstrated that an instructional design can be viewed as an artifact from a DSR perspective and the development of an instructional design can be done according to the phases of a typical DSR research project.

The distinction between descriptive and prescriptive knowledge used in DSR may have highlighted a research gap in the field of information systems. The researcher could not find specific prescriptive guidelines for selection of projects in a PBL driven instructional design.

6. References


