ijecs
open access www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 10, October 2014 Page No. 8510-8518

Design and Implementation of Livestock Data Marts for a Web and Mobile-Based Decision Support System for Smallholder Livestock Keepers: Case Study of Tanzania

Bernard Mussa*[1]*, *Zaipuna Yonah*[1], *Charles Tarimo*[2],**
[1]School of Computational and Communication Science and Engineering,
The Nelson Mandela African Institution of Science and Technology,
Arusha, Tanzania.
[2]College of Engineering and Technology, University of Dar Es Salaam
Dar Es Salaam, Tanzania.

*****Corresponding author's email: ***mussab@nm-aist.ac.tz***

Abstract: *Design of data marts is a fundamental task when preparing data destined for implementation of a Decision Support System (DSS). To answer questions on the underlying users information needs, a data-mart designer is challenged to distill the relevant information from various data sources, enable in-depth data analysis and provide ease of access of information to targeted DSS users. This paper presents the design of data marts for analysis of livestock datasets using dimensional modelling techniques. The designed data marts are then implemented for supporting information and knowledge extraction, leveraging large quantities of livestock data from livestock data source that was identified in the case study environment. Appropriate data mart dimensions and facts were modeled in order to ease data analysis and queries in a role-based information decision support model that was adopted in studied context. Data models based on dimensional modelling (star schema model) are provided and discussed. A comprehensive example showing how a piece of data is loaded from livestock data repository to 'fact' and 'dimension' tables using an open source CloverETL Designer tool is also given. The paper concludes with an overview of the overall detailed schema for the livestock data mart that will serve as a backend engine for On-Line Analytical Processing (OLAP) analysis, reporting, data visualization and information querying via mobile and web access. It is anticipated that the overall DSS once implemented, can be used for improving information delivery, sharing and decision making process to smallholder livestock keepers and livestock experts in Tanzania.*

Keywords: Decision Support System, Data Marts, Dimensional Modelling, Smallholder livestock keepers.

1. Introduction

The growing volume of data in organizations fuels the demand for applications with data warehousing features for in-depth analysis and ease of access to information by targeted users. Data models form the foundation of data warehousing systems since they help to describe how data is to be represented and accessed [1]. In essence, data models form the blue print for the development of databases, which in turn form the backbone of information systems.

For this reason, it is preferred to develop a DSS based on data warehousing techniques. Data marts are the databases that users actually go to for information. An efficient DSS is therefore needed to extract the required user demanded data from data repositories [2].

A data mart is a simple form of a data warehouse that is focused on a single subject (or functional area), such as Sales, Finance or Marketing [3]. The merits of creating data marts include: - users can access regularly desired data easily, generates combined view by a group of users, enhances end-user response time and this is done at a lesser cost than employing a complete data warehouse.

Data marts are built to address the analytical needs of specific sets of users [4]. In this case study, smallholder livestock keepers and livestock experts from Meru District, in northern parts of Tanzania, were surveyed to identify their specific information needs towards development of a DSS that uses data marts.

The main purpose of this work was to develop data marts capable of dealing with the huge amounts of data available at the Meru District Council's Livestock Database System towards valuable and accurate information retrieval for quicker, better quality, and data-driven decisions. Operationally, data from the said data source are extracted and loaded into data marts and presented to users who access them using a variety of tools (i.e. mobile and web access) via the DSS interfaces.

Teste defines a data mart as subject-oriented and dedicated to a specific class of users and it regroups all relevant information for supporting their decisional requirements [5]. Furthermore, creating a data mart involves more than reading columns from a source table to the data mart [6]. Therefore, data mart design process is a critical success factor for being able to answer the underlying business questions that meet the information requirement of the targeted users in supporting their decision making. This paper presents the work done to build and implement adequate data marts for the DSS proposed in [7].

This paper is organized in five (5) sections. The section 2 briefly summarizes related literature on design and implementation of data marts. Methodologies employed in data marts design and implementation processes are discussed in Section 3. Section 4 covers results and discussion of the implementation process of livestock data marts. And in Section 5, the paper ends with a conclusion of the paper on design and implementation of livestock data marts for a web and mobile-based DSS for smallholder livestock keepers.

2. Literature Review

Selecting a data modeling technique for an information system is determined by the objective of the resultant data model. The authors in [1] demonstrate that, dimensional modeling as being the preferred modeling technique for data destined for data warehousing systems. They present data models that ease analysis and queries that are in contrast with entity relationship modeling. For this reason, the dimensional modelling design approach is adopted in the work reported in this paper.

In order to improve analyses and decision making process with regards to the adopted Role-based Information Decision Support (RIDS) model as described in [7], the data mart must be modelled "multidimensionally" as emphasized in [5] so as to produce a design schema that has taken into consideration various roles in relation to available data, information flow and decision making process.

The multidimensional model that is defined in this paper is based on the idea of the "constellation" which is explained in [8], in which data marts are composed of several facts and dimensions. Each dimension is shared between facts and it can be associated with one or more hierarchies, thus facilitating comparisons between several measures/facts [5]. The predominant approach for structuring a data mart is the star schema [9].

A dimensional model is a logical design technique that seeks to make data available to the user in an intuitive framework that is intended to facilitate querying [10]. It captures metadata similar to that in an Entity-Relationship (E-R) model but uses a different structure (i.e. the star schema). The primary components of the star schema are fact tables and dimension tables. A fact table is at the center of the star, and the dimension tables form the points [11].

In general, Data Mart stands for highly-focused version of a Data Warehouse (DW) [12]. Due to the highly focused nature of data in data marts, the approach employed in the reported study had focused on the value of a data mart in order to ensure efficiency when delivering a relevant and complete subset of data residing in a data source to perform a decision support [5]. Consideration should be given to setup costs and development time as these can be significantly reduced when right data models are designed.

Building data marts should focus in addressing the analytical needs of individual users and departments [13]; and like the larger DW, data marts typically contain historical data. Selected data is summarized to a level adequate to meet the intended analytical and information needs for inclusion in the data mart. For example, in our data model design and implementation we have separated the data that are typically not needed by smallholder livestock keepers but that might be of interest to extension officers to ensure relevant information are fed to respective roles in the contextual information flow.

The authors in [14] contend that the purpose of data warehousing is to combine core business and data from other sources in a format that facilitates reporting and decision support [14]. In our study, we have used data from an operational livestock database system available at Meru District Council as our primary data source for the developed data marts using the approach presented in [9,13]. In this case, data for the data mart may come either exclusively from a DW, certain operational systems or both.

Furthermore, it is worth mentioning that related works in data marts design and implementation have all stressed on multidimensional view of data during design process and implementation of data mart powered with information obtained through an Extraction-Transformation-Loading (ETL) process from the data source. ETL tools are pieces of software responsible for the extraction of data from several sources, their cleansing, customization and insertion into a DW/Data Mart [15]. 'Facts' and 'dimensions' tables will facilitate a subject-oriented, time-based analysis [16]. Data marts are usually populated using a top-down approach in which data flow from source systems into data marts [17]. This approach has also been widely adopted in the study reported in this paper.

3. Methodology

The methodology employed in the reported research work was based on RIDS model that is described in [7] for identifying information needs and role in the decision making process in the context of the case study environment. Fact and dimensions tables for the data marts were designed using dimension modelling techniques, and top–down approach was adopted to model resulting tables into a star schema for the data mart implementation. As stated in [3], the major steps in building a data mart are:- Designing, Constructing, Populating, Accessing and Managing. The reported work in this paper discusses designing, constructing, and populating steps.

3.1 Data Mart Design Process

To identify and build data marts the design should be driven by the purpose that each data mart is expected to address [18]. As a consequence, we argue that, the data mart design process must be based on a deep understanding of the expected usage. Thus, the data mart agile development framework in [16] was also adopted to ensure a systematic approach is followed in the design process. Fig. 1 depicts the agile development framework.

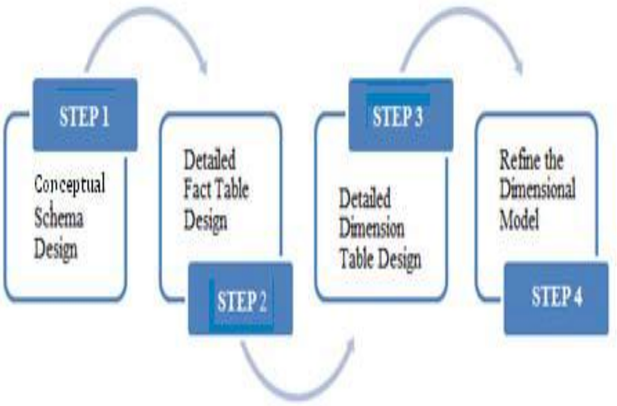
STEP 1
Conceptual Schema Design
Detailed Fact Table Design
STEP 2
STEP 3
Detailed Dimension Table Design
Refine the Dimensional Model
STEP 4

Figure 1: Data mart agile development framework. (Adopted from [16]).

A MySQL database schema model was obtained through a Reverse Engineering process from the Livestock Database System identified in [7] as a source system. A conceptual star schema design was developed on the idea of the multidimensional model (constellation) as defined in [8].We note that, the star schema is the most popular way to build high performance data mart data structures in a relational environment [9]. Moreover, detailed Fact and Dimension tables were designed by using demand-supply driven approach in requirements mapping techniques explained in [7]. The resulting model was implemented using SQL Power-Architect Software and MySQL database engine.

3.2 ETL Processing Tool and Data Loading

An open source, CloverETL Designer Community v.4 was used as an ETL tool for the process of Extracting, Transforming and Loading data from the source system and populating relevant data into the designed fact and dimension tables.

Data marts are usually populated using a top-down approach in which data flow from source systems into an enterprise DW and from there into data marts [17]. With our case study, Meru District Council do not yet have a DW. Therefore, the data marts were populated with data extracted directly from the Livestock Database System (source system). The architectural block diagram in Fig. 2 illustrates the flow of data in the extraction process.

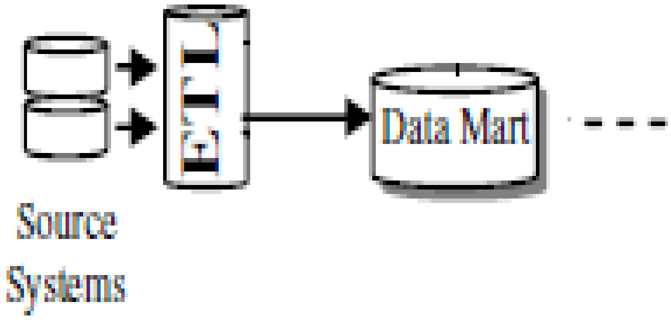

Figure 2: Architectural block diagram.

4. Results and Discussion

4.1. The Conceptual Data Mart Design

A design schema of the data mart is done by collecting all the necessary requirements of the users and identifying the data sources of the collected data. Thereafter, an appropriate subset of data and an analysis and design of the physical and logical structure of the data mart is analyzed and then designed [2]. Once this is done, user requirements are mapped into a conceptual schema of the data mart, that is, a platform-independent, non-ambiguous, comprehensive representation of the facts for decision support that gives a multidimensional picture of the data mart content to both designers and business users.

In contrast to supply-driven approach and demand-driven approach that are described in [19], a mixed demand-supply driven approach was used in this work, from which users' information requirements were mapped into the available data from the identified source system (i.e. Livestock Database System). The source database schema is presented in Fig. 3. Consequently, the Conceptual Design of the data mart that presents a logical structure of the underlying data marts was created as shown in Fig. 4.

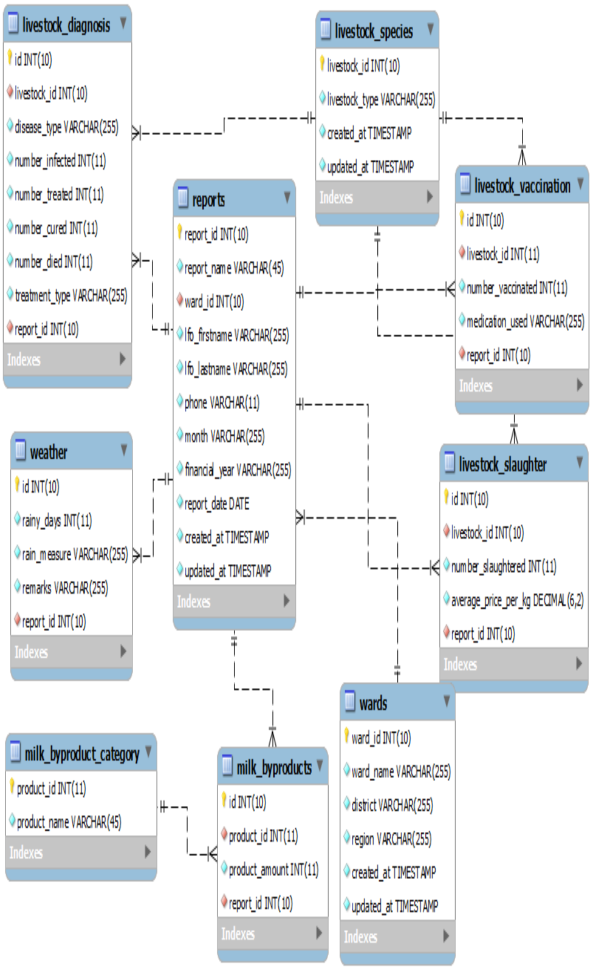

Figure 3: Source Database Schema.

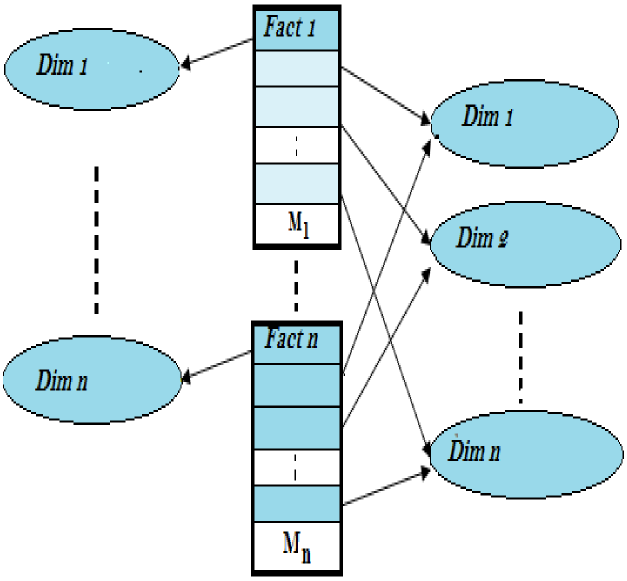

Figure 4: Data mart Conceptual Model.

As a best practice, the adoption of a conceptual model breaks a software design into two distinct but interrelated phases that are largely independent of the features of the

OLAP engine chosen for deployment namely:-logical and physical designs. Creating the physical and the logical construction connected with the data mart gives fast access to the data in data mart [2].

4.2. Logical Data Mart Design

A logical design is a conceptual or an abstract design. It deals only with defining the types of information that is needed. In this paper, we have created an optimized logical schema for the data mart, that is, a set of star schemata. Conveniently, the resulting logical schema can be fine-tuned based on the expected data volume and on the designer's preferences.

A: Facts and dimensions detailed design

With respect to the detailed data mart design, the developed designs involved a four step design process: - 1) Selecting business processes to be modelled, 2) Declaring the data grain, 3) Choosing dimensions, and 4) Identifying facts or measures.

To come up with the detailed design, it is recommended to firstly identify those measures that the user needs. As a result of user interviews and requirement gathering steps reported in [7], business processes were selected and essential data elements needed were identified.

Granularity represents a very important aspect in the analysis of data, as it determines the level of available information [16]. By definition, the granularity or data grain of a fact is the level of detail at which the respective fact is recorded (i.e. the level of detail for the measurement) and made available to the dimensional model. Raw data from the data source (highest level in this case) were the granular level of detail that was selected due to the flexibility in changing to lower level through summarization.

A Fact table is the primary table in a dimensional model where the numerical performance measurements of the business are stored while a Dimension tables are integral companions to a fact table .The dimension tables contain the textual descriptors of the facts [20]. Facts are the numeric metrics of the business. A fact table contains raw numeric items that represent relevant business facts (price, number of infected livestock, livestock product amount and so on.). Dimension tables represent the different ways that data can be organized, such as location, time, and livestock species, etc. For the livestock data mart, dimensions and facts tables were chosen as summarized in Table 1.

Table 1: Facts and Dimension Attributes

S/N	Fact Tables	Dimensions Tables
1	Disease Surveillance Fact Table **Measures/Facts:** • Number infected	Location (Ward) Time(Month) Livestock type(Cattle) Disease Type

2	Market Price Fact Table
	Measure/Facts:
	• Price per Kg
		Location(Ward)
		Time(Month)
		Livestock type (Cattle)

3	Milk Production Fact Table
	Measure/Facts:
	• Product amount
		Location(Ward)
		Time(Month)
		Product Type(Milk)

4	Weather Fact Table **Measure/Facts:** • Rainy days	Location (Ward) Time(Month)

B: Star schema design

As with any other schema design, star schema design is a way to interpret user and business requirements and then model them as data structures [9]. A star schema is a method of organizing information in a data warehouse that enables efficient retrieval of business information [21]. It consists of a collection of tables that are logically related to each other. Using foreign key references, data is organized around a large central table, called the "fact table", which is related to a set of typically smaller tables, called "dimension tables." An example of logical data mart design for a disease surveillance data mart is as presented in Fig. 5.

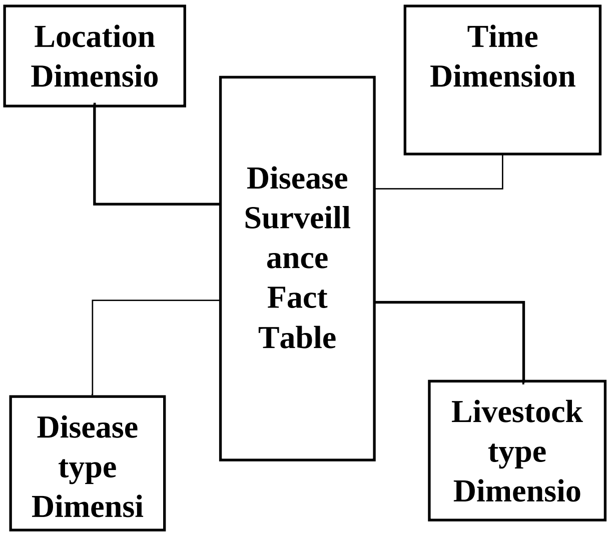

Figure 5: A logical data mart design example (Star Schema Approach).

Star schemas are used in data warehousing as a primary storage mechanism for dimensional data that is to be queried efficiently. The feature of efficiency built in a star schema is important especially as the data volumes that are required to be stored in the DW/Data Mart increase [21].

4.3. Physical Data Mart Design

Data gathered during the logical design phase was converted into a description of the physical database, including tables and constraints. The physical design shows the description that optimizes the placement of the physical database structures to attain the best performance [22]. Due to the nature and types of queries that data mart users usually execute, the data mart database was optimized to perform well for those types of queries. Physical design decisions, such as the type of index or partitioning, have a huge impact on query performance. The livestock data mart tables and constraints were designed as shown in Fig. 6.

4.4. Data Extraction, Transformation and Loading (ETL)

For data marts to serve their purpose of facilitating data analysis, data need to be loaded into them regularly. This is attained by extracting and coping data from one or more operational source systems. In the reported study, the Livestock Database System was identified as the source system from which data was extracted. During extraction, the desired data was identified and extracted into the staging area for further manipulation. This was the first step in the process of getting data into the data mart environment.

After data is extracted, it has to be physically transported to the target system or to an intermediate system for further processing. The transformation involved joining source database tables and re-formatting some of the columns to conform to constraints in the target data mart schema. These transformations are all precursors to loading the data into the physical data mart storage that was implemented in MySQL engine.

The ETL processes described above were all implemented using open source CloveETL Designer Community v4 as depicted in Fig. 7. This procedure was repeated to load all other fact tables as well as the dimension tables as per the required constraints of their schema design

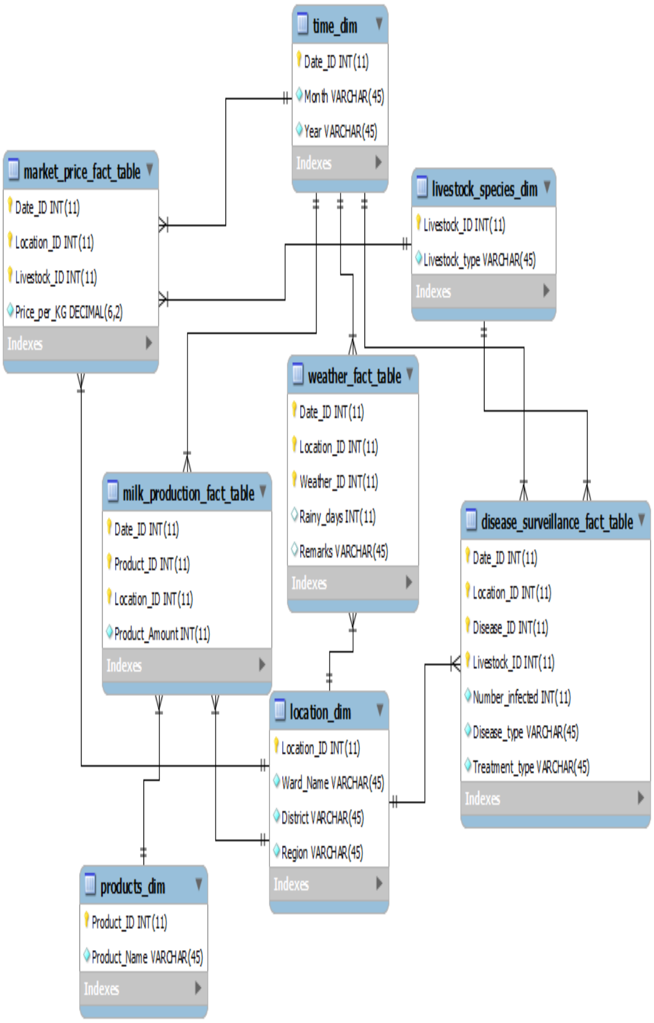

Figure 6: Physical Data Mart Schema.

Data that was extracted and loaded into data marts from the said data source were real historical livestock data obtained from monthly data collected by Meru District Council's Livestock Office spanning three (3) months. Since data marts are the databases that Decision Support Systems query, data marts together with the loaded data serve as back end database for users' access through various tools and interface provided by the DSS environment as illustrated in Fig. 8.

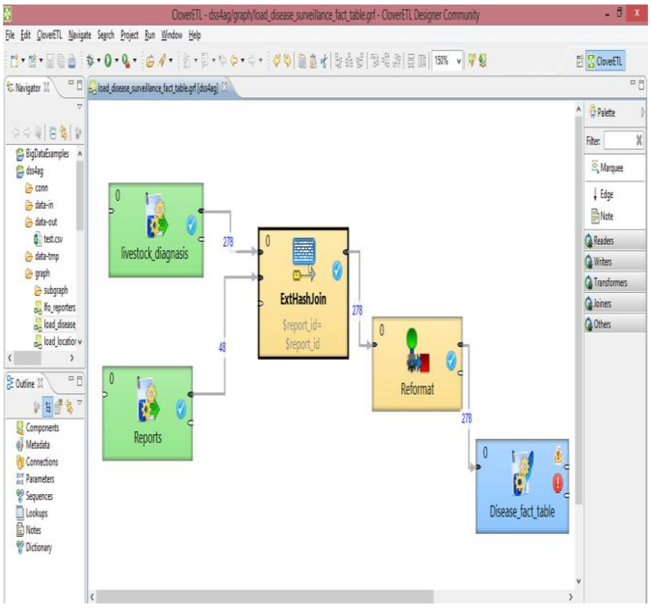

Figure 7: ETL process for loading Disease Surveillance Fact Table.

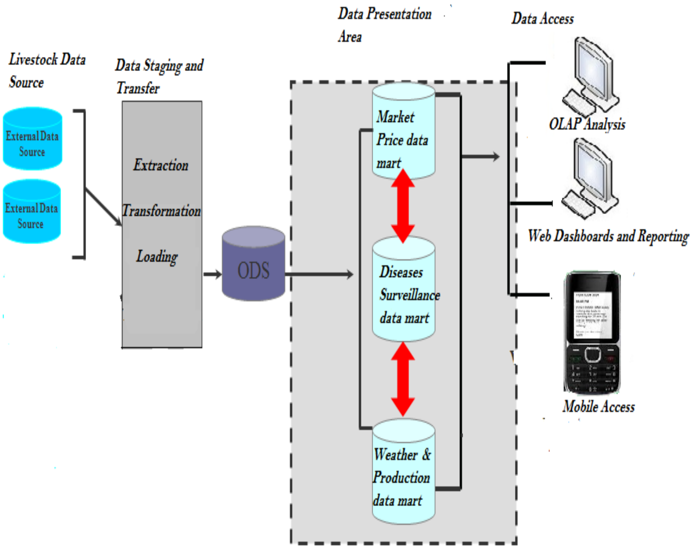

Figure 8: DSS data flow.

5. Conclusion

DSS systems that access operational data are increasingly becoming critical to organizations that wish to exploit operational and other available data to improve quality of decision making, as well as gain critical competitive advantage [10]. In this paper, we report on created data marts that can be used to support in-depth data analysis, efficient reporting and querying of information that can be shared to intended users for decision making support. The driving business factor for the data mart is the need for information and thus the objective of data mart design was easy access to relevant information to smallholder livestock keepers and livestock experts in Tanzania. Its implementation streamlines information delivery for decision support. The models and techniques presented in the paper are by no means a complete set, but they provide a good starting point when new data marts projects are undertaken. The Livestock data marts developed will serve as database for the DSS data visualization module to be developed. The DSS system provides several types of data access, sharing and delivery capabilities (mobile and web) for retrieving and analyzing the data contained in data marts and eventually providing the critical information needed by livestock stakeholders for their decision making.

Acknowledgements

The authors express their appreciation to Nelson Mandela African Institution of Science and Technology (NM-AIST) for providing financial and material support for the reported research undertaking. Support provided by the Meru District Council's Livestock Office is also recognized and acknowledged.

References

[1] Otine, C., Kucel, S. B., & Trojer, L. (2010) Dimensional modeling of HIV data using open source. In *Proceedings of World Academy of Science, Engineering and Technology* (Vol. 62). World Academy of Science, Engineering and Technology.

[2] Paulraj, M., & Sivaprakasam, P. (2012). Functional Behavior Pattern for Data Mart Based on Attribute Relativity. *IJCSI International Journal of Computer Science Issues*, *9*(4).

[3] Dunbar, V. (1997). The Oracle Data Mart Suite Cookbook. *Oracle product documentation, Oracle Corporation*.

[4] Calì A., Lembo D., Lenzerini M., Rosati R. (2003) Source Integration for Data Warehousing. *Multidimensional Databases*, pp 361-392.

[5] Teste, O. (2010). Towards conceptual multidimensional design in decision support systems. *arXiv preprint arXiv:1005.0224*.

[6] Svolba, G., & Austria, S. A. S. (2006). Efficient of a "One-Row-per-Subject" Data Mart Construction for Data Mining. *In Proceedings of the SAS Users Group International 2006 Conference* (pp. 078-31).

[7] Mussa, B., Yonah, Z. & Tarimo, C.. (2014). Towards a Mobile-Based DSS for Smallholder Livestock Keepers: Tanzania as a Case Study.. *International Journal of Computer Science and Information Security*. 12 (8), p54-63.

[8] Pokorný, J., & Sokolowsky, P. (1999). A conceptual modelling perspective for Data Warehouses. In *Electronic Business Engineering* (pp. 665-684). Physica-Verlag HD.

[9] Hobbs, L., & Hillson, S. (2000). *Oracle8i data warehousing*. Elsevier.

[10] Kimball, R. (1997). A dimensional modeling manifesto. *DBMS*, *10*(9), 58-70.

[11] Centers for Medicare &Medicaid Services. (2014). *Dimensional Data Design - Data Mart Life Cycle*. Available: *https://www.cms.gov/Research-Statistics-Data-and-Systems/CMS-Information Technology/DataAdmin/downloads/DimensionalDataDesign.pdf*. Last accessed 18th Aug 2014.

[12] Houari, N., & Far, B. H. (2004, May). An intelligent project lifecycle data mart-based decision support system. In *Electrical and Computer Engineering, 2004. Canadian Conference on* (Vol. 2, pp. 727-730). IEEE.

[13] Bukhbinder, G., Krumenaker, M., & Phillips, A. (2005). Insurance Industry Decision Support: Data Marts, OLAP and Predictive Analytics. In *Casualty Actuarial Society Forum* (pp. 171-197).

[14] Allan, R. G., & May, D. R. (2001). Data Models for a Registrar's Data Mart. *Journal of Data Warehousing*, *6*(3), 38-53.

[15] Vassiliadis, P., Simitsis, A., & Skiadopoulos, S. (2002, November). Conceptual modeling for ETL processes. In *Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP* (pp. 14-21). ACM.

[16] Muntean, M. I., Târnăveanu, D., & Timişoara, (2012). A Multidimensional View Proposal of the Data Collected Through a Questionnaire. Associated Data Mart Deployment Framework. *Database Systems Journal*, *3*(4), 33-46.

[17] Myers, R. (2012). *Institutional Research Data Mart: Student Guide*. Available: *http://www.webgrok.com/rmyers/print/dmadvorg.pdf*. Last accessed 20th Aug 2014.

[18] Miska, M., Gajananan, K., Chung, E., & Predinger, H. (2011). A traffic simulation standard based on data marts. In *Proceedings of the Australasian Transport Research Forum 2011* (pp. 1-11). PATREC.

[19] Battaglia, A., Golfarelli, M., & Rizzi, S. (2011). QBX: a CASE tool for data mart design. *In Advances in Conceptual Modeling. Recent Developments and New Directions* (pp. 358-363). Springer Berlin Heidelberg.

[20] Corey, M. J., & Abbey, M. (1996). Oracle data warehousing. Osborne/McGraw-Hill.

[21] Rausch, N. (2006). Stars and Models: How to Build and Maintain Star Schemas Using SAS® Data Integration Server in SAS® 9. *Proceedings of the Thirty-First SAS User's Group International*.

[22] Stackowiak, Robert, Joseph Rayman, and Rick Greenwald. Oracle data warehousing & business intelligence SO. John Wiley & Sons, 2007.