Use of Date Warehouse in Animal Husbandry and Animal Breeding

R.Reents[1], B.Waurich[2], M. Witt[1]

[1] vit w.V., IT Solutions for Animal Production, Heideweg 1, 27283 Verden, Germany
[2]RBB, Rinderproduktion Berlin-Brandenburg GmbH, Lehniner Str. 9, 14550 Groß Kreutz (Havel), Germany

Abstract

Sources for data in animal husbandry and animal breeding are becoming more heterogeneous. In addition to the typical data from identification, registration, milk recording, AI, and type classification there are 'new' data such as on functional traits and data used to generate information in decision support systems for farmers. The complexity of this data leads to integrated databases which must try to keep historic and current data to generate, for example, across farm benchmarks covering extended time periods. However, the need to summarise and visualize complex information leads to expensive software development if the reporting is programmed in languages such as Java. In some cases not all data from disparate sources can be readily incorporated into one integrated (relational) database system.

vit (IT Solutions for Animal Production) in Germany has undergone a complete computing platform change, from a mainframe based system to a Linux operating system with an Oracle database and software development via Java. A data warehouse (DWH) has been used for reporting and data analysis. In an extract, transform and load (ETL) step a central repository of data is created. The data stored in the DWH are copies from the operational central Oracle systems (such as herd owner information, milk records of the cows, pedigrees of all animals, type classification records, movements etc.). Then data is arranged into hierarchical groups, often called dimensions, and into facts and aggregate facts. From this data source trending reports such as annual and quarterly comparisons are generated.

A comparison of costs for creating reports for farmers or farmer service organisations has shown that the use of a DWH system can be significantly more effective than creating these reports with standard software development tools. Beside the use in reporting a DWH is a very powerful tool to analyse complex data structures to generate new statistics for use in advisory services.

Keywords: data warehouse, big data, analysis

Introduction

vit is a well recognised competent information provider for animal husbandry and breeding. vit is owned by and works for agricultural organisations involved in milk recording, artificial insemination and breeding. From this basis vit operates a central data processing system with

integrated specialist applications and a standardised data inventory. A key feature of this system is the integrated database that contains specific access rights so that competing organisations (like AI centres) can work in one area of the infrastructure and benefit from shared costs and joint data analyses (like genetic evaluation) but can maintain their private data without disclosure to competitors. The information generated from this central database is then provided by different means to the customers, mainly online through attractive and practical browser based applications but also as electronic data transfer via standardized interfaces. Paper is used to disseminate information especially to smaller farms.

Recently the amount of data collected and exchanged has increased exponentially. In large dairy cattle populations like the German one the operative database contains millions of animals. Many applications require historic data and therefore one has to manage e.g. several hundred millions of test day records when analyzing data across generations. A few years ago limited computing capacity made complex analysis of this data very costly. Highly specialized, proprietary software had to be developed to generate information from this complex data structure. Significant programming skills and computer capacities were necessary but often not available. Even small variation in the analysis or report requires a large effort to change the system. Nowadays computing capacity and commercial standard software to handle big data is readily available. vit has addressed this challenge and developed solutions based on Data Warehouse (DWH) technology.

A Data Warehouse is a system that extracts, cleans, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making (Kimball, 2004).

Figure 1 displays the general setup of the vit-DWH. A first step is to define jointly with the customer which data from the overall database is needed in the DWH. These are grouped in data marts that contain data from the same sources (like herd recording data, Herdbook data, sales information, etc.). Then a replicate of the necessary data set is copied into the DWH. New arriving data can be updated based on time definitions (hourly, daily, monthly,…). In this step also additional sources of data, that do not belong to the core business of vit (like AI lab data, information from CRM systems etc.), can be added to the DWH. The accumulation of data in a DWH into a dimensional model with specific data content is called as the ETL (Extract, Transform, Load) process. This can be highly automated to run efficiently in a powerful data centre as vit is operating.

After all necessary data has been defined and accumulated into the DWH an analyzing tool is used to perform the analysis and generate information from this data pool, especially as a decision support system.
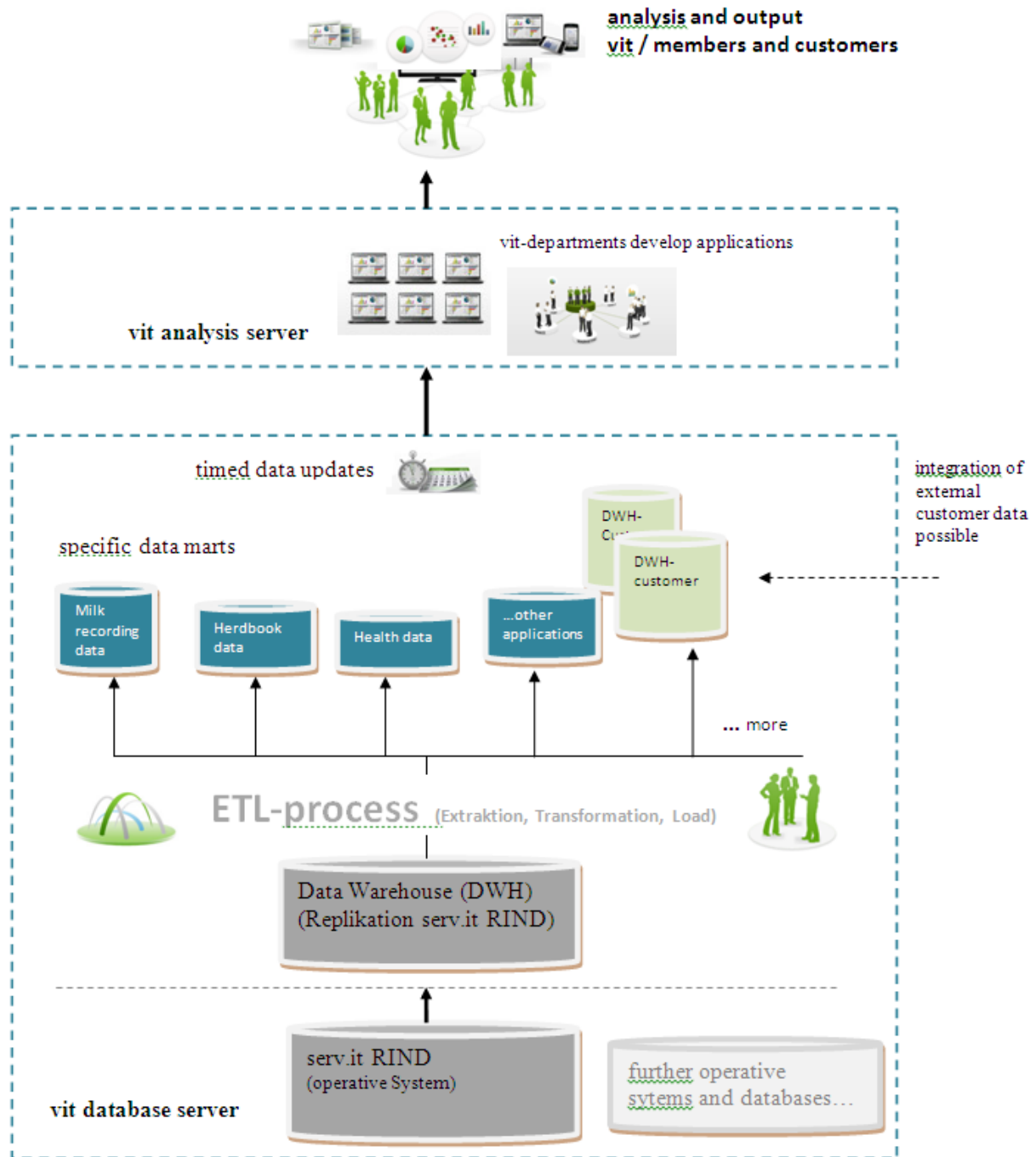
Figure 1. Schematic structure of the data warehouse within vit.

Comparison to traditional ways of analysis

In many cases data queries and analyses using Hibernate Query Language (HQL) or Structured Query Language (SQL) are performed directly on operational databases. The HQL / SQL queries on the operational (nowadays mostly relational) large databases may be very time consuming and always require an intensive examination of the particular queries. In contrast the data that is accumulated in a DWH system can be analysed with much better performance. Data quality is ensured by a well-defined ETL step. The query on operational databases using HQL / SQL requires highly specific knowledge of the user both in the

structure of the database as well as in the query language used. With the use of a DWH these queries can be performed without special HQL / SQL knowledge. Having accumulated all relevant data from different source data avoids different results from parallel reporting on slightly different datasets that is hard to avoid in the traditional way of analyzing complex data.

In an intensive study vit has analysed different DWH tools and has chosen QlikView as a reporting and analysis tool. The data analysis is performed in QlikView in-memory and this enables the user to deliver results nearly without delay (like in an Internet search engine). The in-memory associative search technology combined with an intuitive data exploration toolkit makes it a very powerful system. It can be used in house, in the cloud or offline on a laptop or mobile device from a single user to large global enterprises. Qlik is headquartered in Radnor, Pennsylvania with offices around the world and a network of over 1,500 partners to serve more than 29,000 customers in over 100 countries worldwide (QlikView, 2014).

However, it is a costly tool which requires significant investment in hardware, software licences and in qualified personell that develop the data model and the basic application. This investment has been made by the whole group of vit customers and can now be used for specific applications.

Implementation of services

vit has built up (new) expertise to install and develop customized applications and processes. Beside this the comprehensive knowledge of typical animal breeding and animal husbandry data enable the specialists of vit in cooperation with the customer to develop an optimal solution that is then used individually by the customer with only little support from vit.

Since no one knows better the specific requirements of our animal breeding data than the organizations themselves we enable our customers to work efficiently on their topics and be very flexible when new questions arise. This leaves the analysis of mission critical processes where they belong: in the hands of our users. The DWH is on the way to become a part of everyday data processing along the entire value chain of our animal breeding organizations. In animal breeding and husbandry analytical questions are no longer separate and solely duties of the IT departments but an inherent task of the entire organization.

The advantages of the vit-DWH:

- it assists reporting and analysis needs

- significantly accelerates reporting

- reliable statistics through a uniform data base

- high data security

- user-friendly data processing

- reduction of reporting from other transactional or operational systems

- integration of different sources of data

- identifying and correcting faulty data in the input system

- high performance of queries

- analysis in the DWH does not have negative impact on performance of the operational input system

- short development times

Case Study 1 – Contract herds with additional functional traits

In two regions of Eastern Germany (Mecklenburg-Western Pomerania and Brandenburg) pilot projects have been initiated to record on contract farms additional health and performance data of dairy cattle. A total of 87 herds where selected mainly based on accurate data recording and sufficient genetic variation of production traits. The average herd size was nearly 700 cows in 2013 and represents well the dairy production circumstances in eastern Germany. The aim of this scheme is to collect a large amount of data on functional traits like disease diagnoses, weight at calving, hoof trimming and conformation traits (of all cows) and thus get comprehensive information of all animals in these herds. Substantial progress in genomic evaluation has fostered the broad collection of phenotypes to develop the tools to select animals for these novel traits.

In a first step these new traits were added to the integrated database of vit to use the well-established tools for transmitting this data from on farm management software (where data is captured) to the central database. Key for good data quality is also an early quality check by comparing arriving data with historic data and data from other sources (like AI data, milk recording data etc.). After checking and data cleaning the acquired health data are incorporated into the relational data model of the vit cattle systems. Then this health data is linked to traditional data like calving and performance data. After the ETL step all data is available in one system.

In a next step the report for the herd managers that need appropriate tools for short-, middle- and long-term decisions are produced. Herd management programs serve well for daily decisions, mainly for production traits. The monthly reports of the milk recording organisations allow for across herd benchmarks and are an important source of information for the herd manager. However, for the collected functional traits appropriate reports were not available. But especially here, periodic benchmark-reports serve as good in-herd overviews, but also provide real comparison and benchmarking for between-herd comparisons.
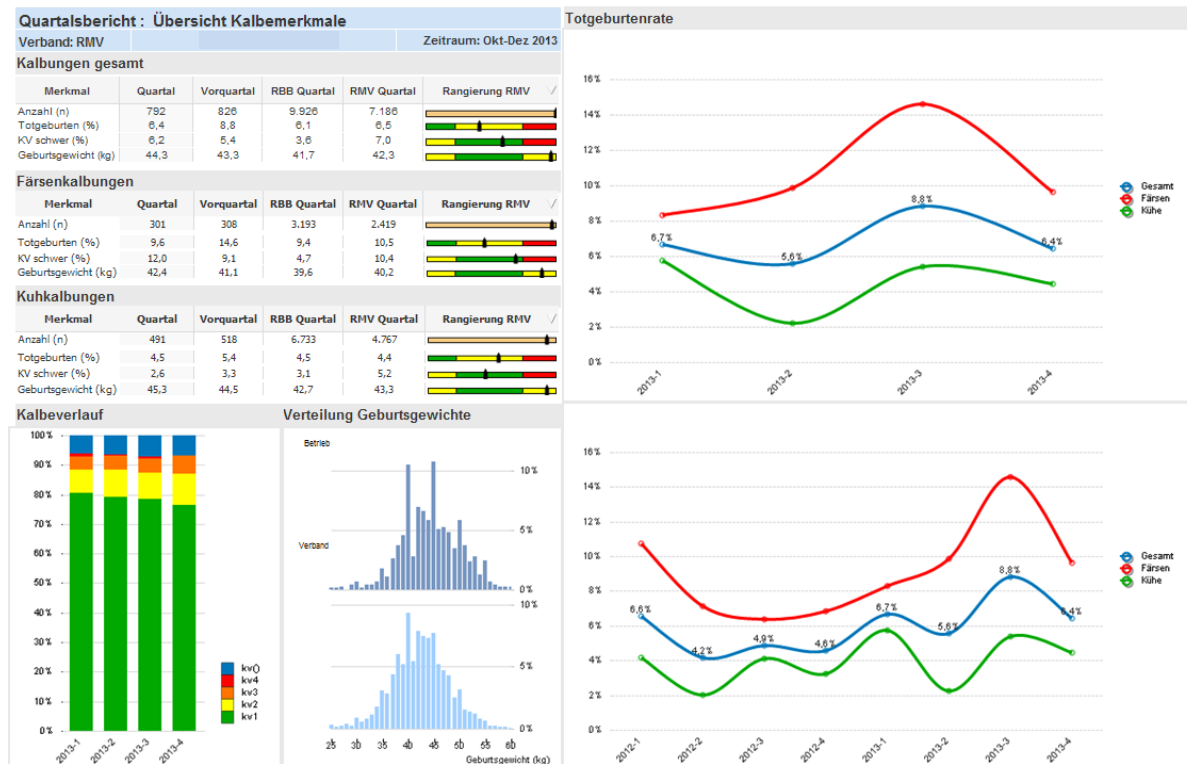
Figure 2. Application of health data and Test herds features.

The printed version of the report is given out quarterly and currently contains six work sheets, which present in graphic and in tabular form the results of calving traits, heifer growth traits, milk production and fertility traits, and extensive health trait analysis. The aim of the report was to hand the herd managers the relevant parameters in an optically attractive layout without overwhelming them with unnecessary data. The feedback of the first routine reports indicates, that these goals have been achieved.

Since health trait documentation guidelines still contain variability due to the veterinarians or herd managers documentation philosophy, interpretation of the between-herd comparisons has to be done carefully.

The data model in the background links each documented performance (or disease, weight, etc) of an animal with a central linkage-table, where each performance can be grouped by the dimension Herd, Date and Animal or a group of them. Whilst single performances work well in this system, special ways had to be found, if a cows calving has to be linked with her calf's birthweight or the weight of a stillborn calf. Where in the first case the offspring has to occur in both the mothers and its own data, in the second case there is a performance by an individual not known by ID. Furthermore, it was difficult to solve questions about contemporary groups e.g. healthy animals in relation to their sick comrades, because disposition/ownership in a herd is just given as two dates between which an animal belonged to a herd. In all these cases, solutions could be implemented, thus proving that the infrastructure can handle complex issues.

Case Study 2 - Milk recording and annual reports

For the different regional milk recording organisations a system for the analysis of the annual results was developed as a replacement for the former mainframe application. Here a variety of statistical analyzes were implemented. The regional organisation can use the 'standard reports' that are similar to the old mainframe implementation but now have the additional possibility to do evaluations on many different levels.
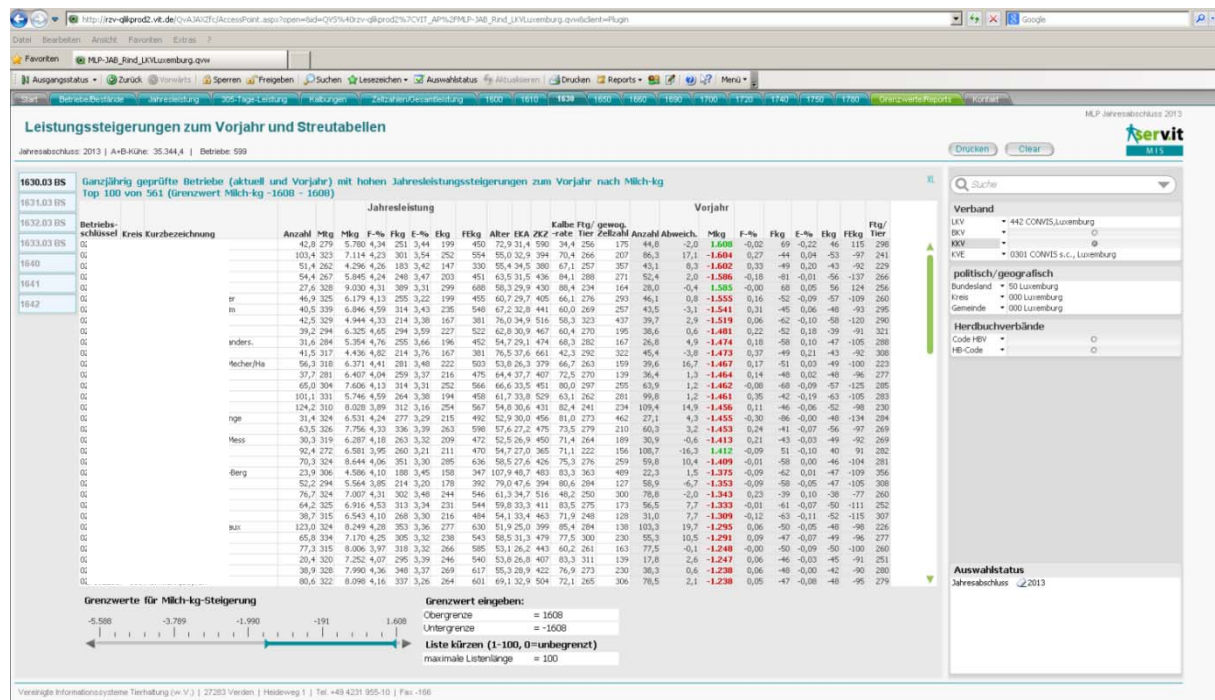

Figure 3. Online-Application "Milk recording and annual accounts" with 1,6 Million A+B Cows, some 15 thousand farms and overall 130 million records.

Case Study 3 - Artificial insemination accounting / billing

Similar to the system for the milk recording organisations vit developed a reporting and analysing system for the AI centers. This gives them a powerful tool for the analysis of their distribution of semen, and analysis of the billing, etc. This makes it possible for the users to carry out independent analysis of their business data anytime and at any place.
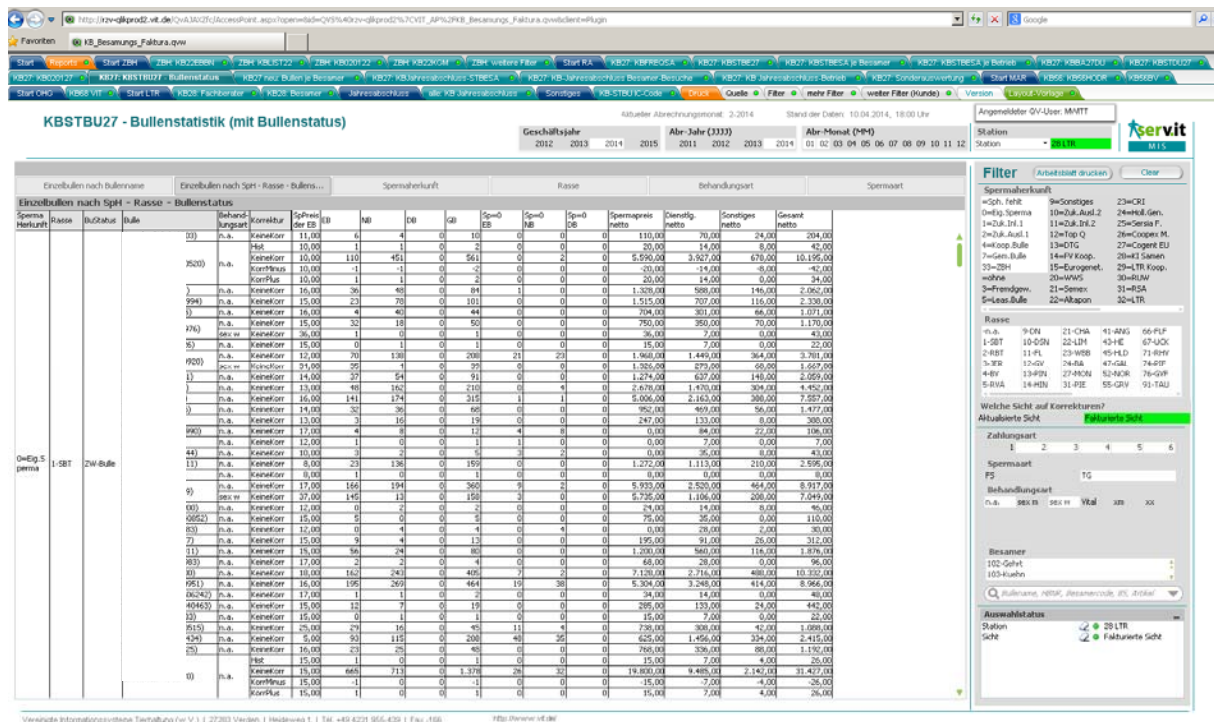
Figure 4. Online-Application "Artificial insemination Billing".

Evaluation of the implementation

The introduction of a DWH system was a high investment with some risk of failure because of its complexity. The scale vit is operating at (being IT provider for a dairy industry with nearly 2 million herdbook Holstein cows) enable this investment to be made and resulted in significant savings compared to a traditional way of programming the reports and analyses tools. Customers have to be well trained to gather the potential benefits from a DWH but more specialized staff at the customers of vit and more technical skills help to implement it on the customer level. So, in general, a very good cost-benefit ratio can be stated given the added flexibility the new system gives.

References

Kimball, R., Caserta, J., 2004. The Data Warehouse and ETL Toolkit – Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Wiley Publishing, Indianapolis.

Qlikview. Qlik Technologies, Inc. 2014, Retrieved from http://investor.qlik.com/index.cfm.