Collecting, integrating, harmonizing and connecting data from dairy farms: the US Dairy Brain project experience

V.E. Cabrera[1], L. Fadul-Pacheco[1], S. Wangen[1], T. da Silva[1], F. Zhang[1], R.H. Fourdraine[2] and J. Mattison[3]

[1]University of Wisconsin-Madison, Department of Animal and Dairy Sciences, 1675 Observatory Dr., 53706, Madison, Wisconsin, USA
[2]Dairy Record Management Systems, 313 Chapanoke Road, Suite 100, 27603, Raleigh, North Carolina, USA
[3]Dairy Herd Improvement Association, PO Box 930399, 53593, Verona, US
Corresponding Author: vcabrera@wisc.edu

Abstract

Modern dairy farms in the US generate vast amounts of data, and with constantly emerging new technologies and implementations, the frequency, diversity, and sheer quantity of these data are increasing. While each source of data can be valuable on its own, the integration of data from different on-farm sources offers a worthwhile opportunity to add significant value to the processes of farm management and decision-making. While this diversity in data collection platforms can be beneficial to the farmer, and to the adoption of automation procedures in general, this diversity also complicates the integration of data from all of these different systems. There is currently no standardization of data organization from a single source (e.g., milking parlor data), meaning that successful integration of data at a large scale requires that each individual source x vendor combination has a unique standardization process to make the data interchangeable and generically available to analysis algorithms.

As part of the University of Wisconsin Dairy Brain project, we are developing an Agricultural Data Hub (AgDH) to do just that. The AgDH is a system which obtains data from on-farm sources and implements a variety of parsing scripts, each one designed to handle the translation of data from one source x vendor combination into a source-specific (but vendor-generic) format. This standardized data structure is then stored and organized in a way that reflects the relationships and interdependencies between different on-farm data sources, facilitating the integration of on-farm data sources, and making the data available for further analysis.

The AgDH is being implemented using an extensible Apache Airflow system of Directed Acyclic Graphs defining a library of workflows or sequence of instructions that orchestrate container-based standardization algorithms and Structured Query Language based data storage along with a data-serving Application Programming Interface endpoint that makes it available to the analytical services further down the value chain.

Keywords: Dairy brain, agricultural data hub, data integration.

Introduction

Dairy farms are data rich but analysis poor because their data streams reside in isolated silos. Insights from data integrated deployed to dairy farms can make large strides in farm efficiency and profitability (Cabrera *et al.*, 2020). Although there are some promising emerging technologies such as Connecterra (2019), JoinData (2020), Idden (2020), the University of Wisconsin-Madison Dairy Brain project (Ferris *et al.*, 2020), among others; the dairy production sector, in general, has been slow to adopt data integrative technologies.

Analytical tools that utilize these integrated data can improve profitability, sustainability, and resilience of farms (Lovarelli *et al.*, 2020). Analyses from integrated data can bring novel insights that are not realized when using only one source of data. These become increasingly important as we move up in the tools' hierarchical level from simple descriptive dashboards to involved predictive simulations to highly sophisticated prescriptive models. Integrated data analysis, also, allows us to envision what would be the unintended consequences of a management change in one area of management to another distant management area, something that is not normally in the radar of the decision makers.

Therefore, as part of the Dairy Brain project, we are developing a framework to connect dairy farm data from various sources and make the data interchangeable and integratable in preparation for downstream analysis, the Agricultural Data Hub (AgDH; Ferris *et al.*, 2020).

The Dairy Brain Agricultural Data Hub (AgDH)

The AgDH collects, cleans, and integrates dairy farm data into a centralized data hub, and makes them accessible to the Dairy Brain analytical modules. More specifically, it involves 5 critical steps: (1) Accessing, (2) Decoding, (3) Cleaning, (4) Homogenization, and (5) Integration (Figure 1).

In brief, data from different sources and of different types (e.g., feed, milk, and health) are collected on-site from standard outputs from different farm software packages and uploaded periodically into the AgDH using minimal software installation on a farm computer (1). Data are then parsed and loaded into a database in a relatively intact native format (2). Next, data are cleaned by verifying validity and duplication (3). Following, same data types from different software are transformed and homogenized to a common format (4). Homogeneous data are then integrated into a data warehouse (5) where they can be accessed via a secure and authenticated web Application Programming Interface (API).

Accessing

The most advanced software components could be cloud-based and offer access to data via API's, but this is still rare in dairy production systems. More often, the data is collected by different systems produced by different companies, and is only available by accessing files stored on a local computer located at the farm. A client process installed in a computer at the farm detects new data files from the systems and triggers the transfer of those files via the internet into a centralized set of servers that host the reminder of the AgDH functionality.

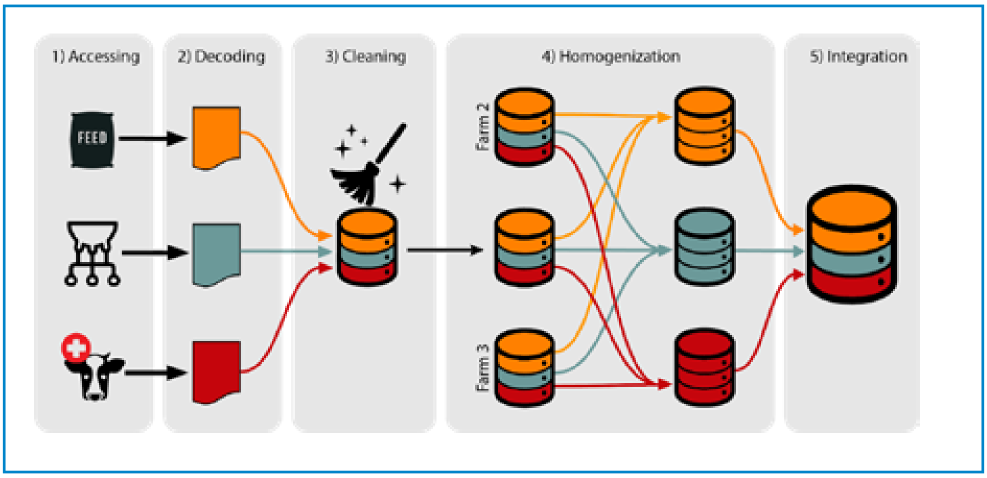

Figure 1. The Agricultural Data Hub (AgDH) framework to integrate dairy farm data.

Decoding

Accessible data are extracted from its raw source and transferred to a more readily available format for easier processing, i.e., ingested. On-farm data sources tend to vary greatly on their level of machine readability due to the fact that existing farm software can output data in many diverse formats depending on vendors, software, and farms, creating a large matrix of data readability type combinations that need to be handled by the ingestion process. Thus, unique routines need to be developed for each data source to handle its peculiarities, which is labor-intense and time consuming. These onerous tasks are alleviated by a modular framework with a flexible approach for parsing scripts that extract and ingest the data. Although there is still a need for custom language for each new software, our framework allows scripts to be combined to assemble a custom ingest pipeline for each farm.

Cleaning

Data used by the system have to meet certain quality standards. Quality ensures the integrity and tractability of the data. Integrity signifies the validity of the data. The first step to guarantee integrity is enforcing data type on the data input replacing data that do not conform with the predefined data type with a null value. Then, exception handling is used to harmonize different null representations that might exist in different systems. Then, a process checks validity of the values by, first, comparing the values against a list of valid entries and then, by applying logic filters such as hard boundary checks, reasonable physical bounds, or mathematical reasoning to identify inconsistent observations. Data deemed invalid are handled with care and transparency following a moldable logic decision tree to determine the level of invalidation (observation, row, or dataset) and documenting the process for human review.

Tractability ensures that any data element and derivative data products can be traced back to its original source and recreated. This is accomplished by capturing every step detail in the data transformation pathway by following a data lineage or data provenance approach and recurring to the use of metadata documents that describe the original source and form of the data. The data lineage begins identifying the farm, the software details, and the format of the data source. Each iteration of the ingestion process receives a unique identifier, which is attached to each entry made or updated from that iteration together with any logs or messaging from the ingestion iteration. These allow to link each data element to a single instance configuration and therefore recall the full ingest process to its original source.

Homogenization

Data follows a horizontal integration that involves identifying and extracting data commonalities among different software that record the same type of data. Hence, it requires standardization of units, terminology, types of measures, intervals, and other relevant details among input data. This process stores and serves standardized data for each farm regardless of the farm specific data collection system. Homogenization is challenging and complex as dairy data is highly heterogeneous and is collected by a number of dissimilar systems.

Integration

Data from different sources are connected following a process of record linkage that involves developing a pipeline to map records from one data source with those from other sources. Since each record from each system does not usually have a global unique key that links the record to a single organism, data connection among systems is complex. Hence, connecting data relies on a combination of linking variables.

With large differences by countries, regions, and even farms, dairy animals are issued a unique identifiable number. In some places, this number could comply with an official government issued number that could or could not be used by the software recording systems of the farm. Most normally, individual animal identifiers in a farm are multiple and likely inconsistently used among data collection systems. Therefore entity match is a critical step for data connections.

Another issue arises from the fact that data on dairy farms resides at different levels of aggregation. Dairy farms have data collected at the individual animal level such as daily milking, data at the group level such as pen feed consumption, and at the herd level such as milk bulk tank composition. Thus, data relationships need to be inferred using yet other data sources. For example, the details of a feeding event from the feeding recording software can be linked to a pen of animals from the event tracking software by using unique animal identifiers and pen allocation. This variation in data available forces aggregated analyses such as averages reducing the tractability of individual animal information and subsequent precision.

Automation and Deployment of the AgDH

We rely on a workflow automation system called Apache Airflow (Airflow, 2020), in which we define a library of workflows or sequence of instructions defined as Directed Acyclic Graphs (DAG) that orchestrate container-based standardization algorithms and Structured Query Language (SQL) based data storage. These DAGs are well suited for diverse data, as the ones encountered on dairy farms, as they can be flexible, dynamic, and modular. All previously defined steps of the AgDH workflow are controlled by DAGs. The automated system retries failed actions and generates alerts and messages if issues persist. The vision is to make the integrated data accessible through API endpoints hosted by the AgDH service. This service will be accessed via secure https connections, which will retrieve JavaScript Object Notation (JSON), Comma Separated Values (CSV), or eXtensible Markup Language (XLM) outputs. The first and main data consumer will be the analytical services of the Dairy Brain, but the system will be prepared and open to serve other research, industry, or consultant groups.

Integrated Decision Support Systems

The vision for the Dairy Brain is a real-time analytical engine capable of performing longitudinal historical analyses and forecasting the future from past information in a continuous loop grounded on data provided by the AgDH. We categorize our models as those descriptive like summary dashboards that show the current situation and might include some simple calculations. A number of our tools will include predictive or forecasting capabilities and will be updated continuously. The most advanced models are conceptualized as prescriptive tools that will provide suggestions, mostly from optimizations, of the best course of action.

Acknowledgement

This project is supported by the Food and Agriculture Cyberinformatics and Tools grant no. 2019-68017-29935/project accession no. 1019780 from the USDA National Institute of Food and Agriculture.

List of references

Airflow, 2020. The Apache Software Foundation. *http://airflow.apache.org/* (accessed 5 November 2020).

Cabrera, V.E., J.A. Barrientos, H. Delgado and L. Fadul-Pacheco. 2020. Real-time continuous decision making using big data on dairy farms. Journal of Dairy Science 103:3856–3866.

Connecterra, 2019. Connecterra. *https://www.connecterra.io/* (accessed 14 September 2020).

Ferris, M.C., Christensen, A., Wangen, S.R., 2020. Symposium review: Dairy Brain—Informing decisions on dairy farms using data analytics. J. Dairy Sci. 103, 3874–3881.

Idden, 2020. International Dairy Data Exchange Network. https://www.idden.org (accessed 4 November 2020).

JoinData, 2020. Join Data. *https://join-data.nl/en/* (accessed 4 November 2020).

Lovarelli, D., Bacenetti, J., Guarino, M., 2020. A review on dairy cattle farming: Is precision livestock farming the compromise for an environmental, economic and social sustainable production? J. Clean. Prod. 262, 121409.