359. From raw sensor and automated data to genetic evaluation and validation in the cloud

D. Schokker[1*], I.N. Athanasiadis[2], M. Poppe[1], J. ten Napel[1], C. Kamphuis[1] and R.F. Veerkamp[1]

[1]*Wageningen University & Research, P.O. Box 338, 6700 AH Wageningen, the Netherlands;* [2]*Data Competence Centre, Wageningen University & Research, 6708PB Wageningen, the Netherlands; dirkjan.schokker@wur.nl*

Abstract

Precision livestock farming tools, such as sensor and automation techniques, enable animal breeders to define novel traits. New challenges are data storage and handling and the development of genetic evaluation. Traditionally, the development of genetic evaluation for novel traits requires several steps, i.e. data curation, trait definition, variance component estimation (ASReml), genetic evaluation (MiXBLUP), and validation of the estimated breeding values (EBVs), requiring many iterations to optimize the genetic evaluations. We combined these steps in a cloud solution to make this entire process more efficient. An experiment was run with 1,782,373,113 daily milk yield records of 1,120,550 cows, all the way from data curation to trait definition and validation of the EBV. The resulting wall-time, i.e. elapsed actual time from start to finish of the entire process, was ~23 hours. The flexible cloud solution can be easily modified or adapted to develop novel traits.

Introduction

The increasing use of sensor and automation technology in livestock farming, for example, to monitor animal health and welfare in cows (Smith *et al.* 2006; Matthews *et al.* 2016; Ouweltjes *et al.* 2021), enable the development of new phenotypes and (complex) traits. The data these sensors generate can be both unstructured non-relational data, including camera or video images, and (non-standardized) structured data, like relational databases (Structured Query Language-SQL). Moreover, sensors are recording in real-time and generate large volumes of data, particularly when recording is done over a long period of time. Big data lakes have emerged as cloud infrastructures that effectively manage these data types, where it is needed to store, manage, access, and process large volumes of a wide range of structured, semi-structured, or unstructured data sources in raw format. In addition, different tools and resources are available in the cloud environment that can interact with such a data lake storage. Such tools and resources include platforms to transform, filter, and structure the data sources, making the cloud environment in combination with a data lake storage both flexible and scalable. Data lake storage can utilise the potential of large volumes of sensor data (Schokker *et al.* 2020). In animal breeding programmes, the challenge lies in utilizing these (novel sensor) data sources to identify the best parents for the next generation. Going from raw (sensor) data to genetic evaluation involves many steps. These pre-processing steps include reading in the data, followed by filtering, transforming, joining and aggregating the data, and writing the output (Gengler 2019). The genetic evaluation includes steps like generating subsets of the data to estimate genetic parameters, and predict breeding values (EBV) and their reliabilities. These downstream steps are performed with different software packages, often in different computing infrastructures. Also, after the final EBV validation, the whole process often has to start again with model adjustments or different data curation. In this study, we combined the entire process in a single cloud environment, linking all the different steps from pre-processing up to the genetic evaluation. As a feasibility experiment, we applied the pipeline to 1,782,373,113 daily milk yield records of 1,120,550 cows to define and evaluate a resilience trait.

R.F. Veerkamp and Y. de Haas (eds)

DOI: 10.3920/978-90-8686-940-4_359, © D. Schokker *et al.* 2022

Materials & methods

Design of cloud environment. We designed and built a custom-made cloud solution that suited our needs to perform all required steps to get from raw (sensor) data to genetic evaluation in a single cloud environment (Figure 1). The focus of this paper is on the infrastructure of this cloud solution.

Within the cloud-environment we employed a BLOB (binary large object) storage because this has tiered flexible storage and scale-up on performance-driven computing. This BLOB storage was linked to other data analytics resources, i.e. Databricks and our customised Docker containers for variance component estimation, genetic evaluation, and validation of the estimated breeding values. Databricks is an Apache Spark-based analytics service designed for data science and data engineering (Zaharia *et al.* 2016), and therefore, well-equipped to handle the nearly two billion milk yield records. Within the Databricks platform, we generated a pre-processing script for the raw (milk yield) data. Such an Extract, Transform, and Load (ETL) procedure is crucial for an easy and flexible scale-up of the process. This whole ETL procedure was written in Python (v3.7.3) and uses Apache Spark (v3.1.1) in the background for efficiency. For the subsequent genetic evaluation, we have used container technology, i.e. Docker.

Experiment. We intended to build the cloud solution for an existing genetic evaluation, so we used the dairy cattle dataset described by Poppe *et al.* (Poppe *et al.* 2020) as a case study. This dataset contained 1,782,373,113 milk yield records from 1,120,550 cows. Additionally, we used the pedigree and date of birth for those animals, which were necessary to curate a dataset for breeding value estimation. We calculated a potential resilience indicator trait from the milk yield records: natural log of the variance of deviations from the lactation curve, as defined by Poppe *et al.* (Poppe *et al.* 2020). This resilience indicator described the fluctuations in a frequently measured trait, here milk yield. This highly frequent sensor and automation data correspond to the developments in data collection. The ETL procedure was set-up in such a way to easily modify the criteria for filtering the data, including: (1) the milking system (automated or conventional); (2) parity (1, 2, >3); (3) breed; and (4) breed percentage.

Pre-processing. The whole pre-processing, i.e. ETL procedure, was performed within Databricks (https://azure.microsoft.com/en-us/services/databricks/), which we deployed on a computer cluster. We set the cluster with the following specification: the driver and workers all had 56 GB Memory and 16 cores, where

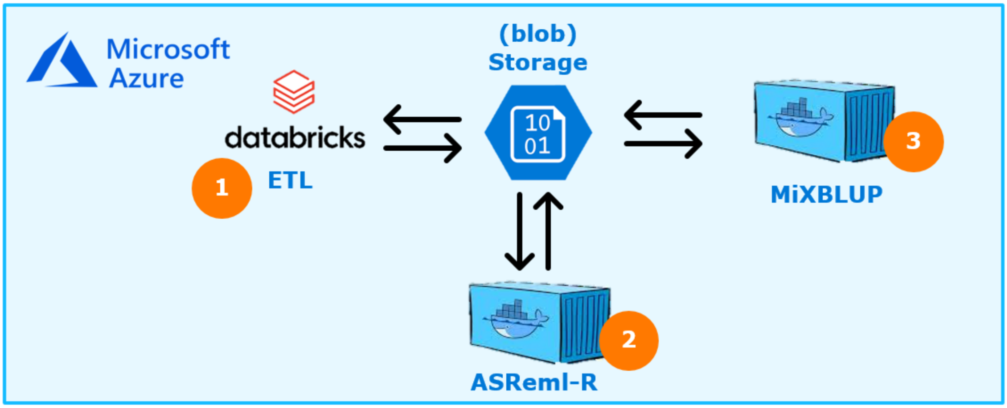

Figure 1. Schematic overview of the infrastructure and resources used in our cloud environment, i.e. Microsoft Azure. Four different modules could interact with the central BLOB (binary large object) storage: (1) pre-processing via Databricks; (2) Docker container with ASReml-R software; and (3) Docker container with MiXBLUP software.

the number of workers may vary from a minimum of 2 to a maximum of 8. Applying the data editing steps using the ETL procedure resulted in 352,871 cows with a resilience trait for at least one of the three lactations in the experiment.

Estimation of genetic parameters. For the estimation of the variance components, first a dedicated Docker (https://www.docker.com/) container was setup within our environment that could communicate via the central (blob) storage. This first Docker container, with Linux operating system, had installed ASReml-R (v4.1.0.110) (Butler *et al.* 2017), and for calculations we used a virtual machine with 2 GPU and 4 CPU cores with 112 GB Memory. Variance estimation is known to be computational intensive, therefore the pre-processed dataset, i.e. 352,871 cows, was split into five random subsets. We obtained five estimates for the heritability and genetic correlations among the three lactations.

Genetic evaluation. In the second Docker container, also with Linux operating system, we installed MiXBLUP v2.2 (ten Napel *et al.* 2017). For the calculations, we used a virtual machine with 4 CPU cores and 16 GB Memory. The model used in MiXBLUP was the same as the model used in ASReml-R. Now the full dataset, i.e. 1,120,500 cows, could be used for the final EBV calculation. From the full pedigree, we build a tailored pedigree starting with the cows present in the pre-processed data and adding five generations of male and female ancestors. We analysed the data using a statistical model with herd-year-season and lactation length as fixed effects, and cow as a random effect (Poppe *et al.* 2020).

Results

For the pre-processing step (ETL procedure) the parameters were set to include automated milking system, parities 1, 2, >3, and the breed Holstein-Friesian with a breed percentage of >75%. The wall-time, the elapsed time from the beginning to the end of the entire process, was approximately 23 hours. Table 1 shows the wall-time for each specific step.

Discussion

We demonstrated that by using a cloud solution we can go from (raw) sensor data to a trait of interest and perform the accompanying genetic evaluation and validation in approximately 23 hours, when processing a very large dataset of more than 1.7 billion milk records from more than 1,000,000 cows. Our results are promising, especially because the expanded use of sensors in the livestock sector will generate an ever-increasing volume of data, either unstructured or structured. This procedure will allow quick alterations in the pre-processing of the data to be tested and validated, therefore, optimizing the time to the final implementation of genetic evaluations using sensor data. The gain is in the generic tools for the data pre-processing and then the 'easy push of a button' for variance component estimation and breeding value estimation, something that often requires a lot of manual work. Our cloud solution enables scalability and flexibility in the cluster computing environment. Examples for scalability include upscaling and downscaling of IT requirements when needed. The flexibility allows easily sharing data, scripts, and results with your collaborators. Implementing the genetic evaluation and validation software in our cloud infrastructure, while avoiding data transferring from one system to another was challenging. Moreover, due

Table 1. Wall-time for each step.

Step	Wall-time (h)
Pre-processing	6
Estimation of genetic parameters	16.5
Genetic evaluation	0.5

to the nature of the cloud infrastructure, it becomes possible to run various genetic evaluations in parallel. For example, estimating breeding values with different models and calculating approximate reliabilities and yield deviations. Another possibility is to run the pre-processing procedure on different clusters, simultaneously, where small but relevant changes in the script can change the outcome of the feature of interest, e.g. specifying the number of parities to include, or what genotype to focus on, or specifying a period in time. These proposed workflows could decrease the wall-time, thus making this process more efficient.

Funding

This study was financially supported by the Dutch Ministry of Economic Affairs (TKI Agri and Food project 16022) and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin.

References

Butler D.G., Cullis B.R., Gilmour A.R., Gogel B.G. and Thompson R. (2017) ASReml-R Reference Manual Version 4, pp. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Gengler N. (2019) J Dairy Sci 102(6): 5756-5763. https://doi.org/10.3168/jds.2018-15711

Matthews S.G., Miller A.L., Clapp J., Plotz T. and Kyriazakis I. (2016) Veterinary journal 217(43-51. https://doi.org/10.1016/j.tvjl.2016.09.005

Ouweltjes W., Spoelstra M., Ducro B., De Haas Y. and Kamphuis C. (2021) J Dairy Sci 104(11): 11759-11769. https://doi.org/10.3168/jds.2021-20413

Poppe M., Veerkamp R.F., Van Pelt M.L. and Mulder H.A. (2020) J Dairy Sci 103(2): 1667-1684. https://doi.org/10.3168/jds.2019-17290

Schokker D., Athanasiadis I.N., Visser B., Veerkamp R.F. and Kamphuis C. (2020) Animal 14(11): 2397-2403. https://doi.org/10.1017/S175173112000155X

Smith K., Martinez A., Craddolph R., Erickson H., Andresen D. *et al.* (2006) An integrated cattle health monitoring system, pp. 4659-4662 in *Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, New York, NY, USA.

Ten Napel J., Vandenplas J., Lidauer M., Stranden I., Taskinen M. *et al.* (2017) MiXBLUP: A user-friendly softwarevfor large genetic evaluation systems, pp.

Zaharia M., Xin R.S., Wendell P., Das T., Armbrust M. *et al.* (2016) Communications of the ACM 59(11): 56-65.