J. Dairy Sci. 105:9792–9798
https://doi.org/10.3168/jds.2022-22113

Rapid turnover of sensor data to genetic evaluation for dairy cows in the cloud

D. Schokker,[1]* M. Poppe,[1] J. ten Napel,[1] I. N. Athanasiadis,[2] C. Kamphuis,[1] and R. F. Veerkamp[1]
[1]Animal Breeding and Genomics, Wageningen University & Research, Wageningen, the Netherlands 6708 PB
[2]Data Competence Centre, Wageningen University & Research, Wageningen, the Netherlands 6708 PB

ABSTRACT

More and more sensor and automation data are available that enable animal breeders to define novel traits. However, sensor and automation data are often frequently measured differently (e.g., milk yield and different milk components are continuously measured during each milking). These differences are challenging animal breeders to define traits and use the most appropriate analytical models for genetic evaluation and breeding values. Traditionally, the process from raw data to breeding value estimations involves several steps: data curation, trait definition, variance component estimation, genetic evaluation, and validation of the estimated breeding values (EBV). All these steps often take many iterations and several research projects to optimize the final genetic evaluations. To make this entire process—from raw data to validated EBV—more efficient, we combined all these steps in a cloud environment that allows for faster processing and a faster data distribution time. We used real data (including 1,782,373,113 daily milk-yield records of 1,120,550 dairy cows) and a real trait (a resilience trait based on the deviations from expected milk yields) to demonstrate the functioning of this cloud environment. The daily milk-yield records were incorporated into our cloud solution, in which we have set up central binary large object storage. Subsequent steps were all performed in the cloud. The data set was preprocessed in approximately 6 h to obtain the resilience indicator for 352,871 cows in the first 3 lactations. Estimation of genetic parameters (heritabilities and genetic correlations) was performed by splitting the data into 5 subsets in ASReml, and prediction of subsequent EBV was performed on the entire data set using MiXBLUP. Together with the validation of breeding values, this process encompassed 16.5 h. By combining the different steps from preprocessing sensor data to genetic evaluation of new traits in one cloud environment, we generated EBV and validation plots in approximately 1 working day. Moreover, our setup is a flexible design and can be adapted easily to test new, longitudinal sensor-driven traits and compare the performance of these new traits to previous ones.
Key words: dairy cows, sensor data, cloud solution, genetic evaluation

INTRODUCTION

The increasing use of sensor and automation technologies in livestock farming enables researchers to define new traits that are linked, for example, to animal health and welfare in cows (Smith et al., 2006; Matthews et al., 2016; Ouweltjes et al., 2021). In animal breeding programs, the challenge lies in using these data sources in breeding value estimations to identify the best parents for the next generation. To develop breeding value estimations for a novel trait requires not only estimation of variance components and predictions to set up the genetic evaluation, but also it includes the processes ranging from loading and preprocessing data to optimizing the statistical model.

The initial step in generating new traits involves storage and handling sensor and automation data. The nature of these data can be both unstructured nonrelational data, including camera or video images, as well as nonstandardized structured data, such as structured query language (SQL). Moreover, these data are often recorded in real time and can generate large volumes of data, particularly when recording over a long period of time. To accommodate these heterogeneous and often voluminous data, data lake storage has been shown to be effective (Schokker et al., 2020). A data lake is a structure to store, manage, access, and process large volumes of a wide range of structured, semistructured, or unstructured data sources in raw format. The adoption of data lakes is often observed in the cloud environment. In addition, such cloud environments also contain tools to preprocess the data, such as reading in the data, and the following preprocessing steps such as filtering, transforming, joining, aggregating, and

Received March 24, 2022.
Accepted August 6, 2022.
*Corresponding author: dirkjan.schokker@wur.nl

writing the output (Gengler, 2019). Such preprocessing is necessary to prepare the raw data for subsequent analyses of variance component estimation and genetic evaluation. These analyses include steps that involve estimating genetic parameters for the trait of interest, which are in turn required for the subsequent genetic evaluation. Because estimating genetic parameters is demanding computationally, this process often includes generating subsets of the data for efficient analysis, as well as EBV and their respective reliabilities. These different analyses are performed with specific software packages, sometimes in different computing environments. In previous work, we reported on the computational time aspects of our cloud solution (Schokker et al., 2022). The objective of the current study was to develop and optimize further the cloud infrastructure for associated genetic software tools [ASReml (Butler et al., 2017) and MiXBLUP], and to test the functioning of the cloud infrastructure in estimating and validating the generated EBV using real (sensor) data from dairy cattle. The results of this test are used in our study to discuss the advantages and disadvantages of our cloud-based approach in generating and validating EBV.

MATERIALS AND METHODS

No animals were used in this study, and ethical approval for the use of animals was thus deemed unnecessary.

Data Infrastructure

All analytical tools and software were implemented within the cloud environment of Microsoft Azure (https://azure.microsoft.com/en-us/), as depicted in Figure 1. Central binary large object (**BLOB**) storage was used and linked to Databricks (https://databricks.com/) and customized Docker containers (https://www.docker.com/). Databricks is an Apache Spark-based analytics service designed for data science and engineering (Zaharia et al., 2010). Databricks was used to extract, transform, and load (**ETL**) the raw data. We wrote a custom Python (v3.7.3) script and used Apache Spark (v3.1.1) for this ETL procedure. The entire aforementioned ETL procedure was performed within Databricks, on which we deployed a cluster. A cluster is a set of computation resources and configurations on which you run your workload. We set the cluster with the following specifications: the driver and workers all had 56 gigabytes of memory and 16 cores, and the number of workers was allowed to vary between 2 and 8. For the subsequent genetic analyses, we installed software on customized Docker containers. This involved setting up the first dedicated Docker container, with Linux as the operating system, that could communicate with the BLOB storage to estimate the variance components. This Docker had ASReml-R (v4.1.0.110) installed, and for calculations we used a virtual machine with 4 central processing unit cores and 112 gigabytes of memory, and 2 graphics processing units (GPU). The second dedicated Docker container, also with Linux as the operating system, had MiXBLUP (v2.2) installed for genetic evaluation and validation of the EBV. For calculations, we used a virtual machine with 4 central processing unit cores and 16 gigabytes of memory.

Data Used

The same data set as described by Poppe et al. (2020) was used, containing 1,782,373,113 milk-yield records from 1,120,550 cows. These milk records were either obtained by automatic milking systems (**AMS**) or conventional milking systems (**CMS**). The accompanying pedigree data had records that went back to the year 1919.

Preprocessing Data and Defining Resilience Parameters

Based on data-editing scripts and snippets from the AWK (shell command) language and R developed by Poppe et al. (2020), we generated a custom-made Python script (as a Jupyter notebook) to preprocess the data by following an ETL procedure. By implementing these various scripts from Poppe et al. (2020) into one common language—in our case, Python (v3.7.3)—we ensured flexibility and interoperability for interacting with Apache Spark (v3.1.1). These newly developed scripts were deployed within a notebook in the Microsoft Azure Databricks environment. In this notebook, different metadata were joined, including the milk records, pedigree, and birth dates. To easily modify the criteria for filtering the data, we made our scripts include several parameters that had to be declared before running the script. These parameters included (1) the milking system (AMS or CMS), (2) parity (1, 2, >3), (3) breed, and (4) breed percentage. In our demonstration case, we set these parameters to (1) AMS, (2) parities 1 through 3, (3) Holstein-Friesian, and (4) breed percentage $>$ 75%. With these parameter settings, we performed several filtering steps of these records and aggregated the data to specific traits. This ETL procedure consisted of 4 aggregation steps: first, calculating the average milk yield per day; second, fitting a rolling average to the lactation curve, where we set the window size for the rolling average to −10 and +10 d; third, calculating

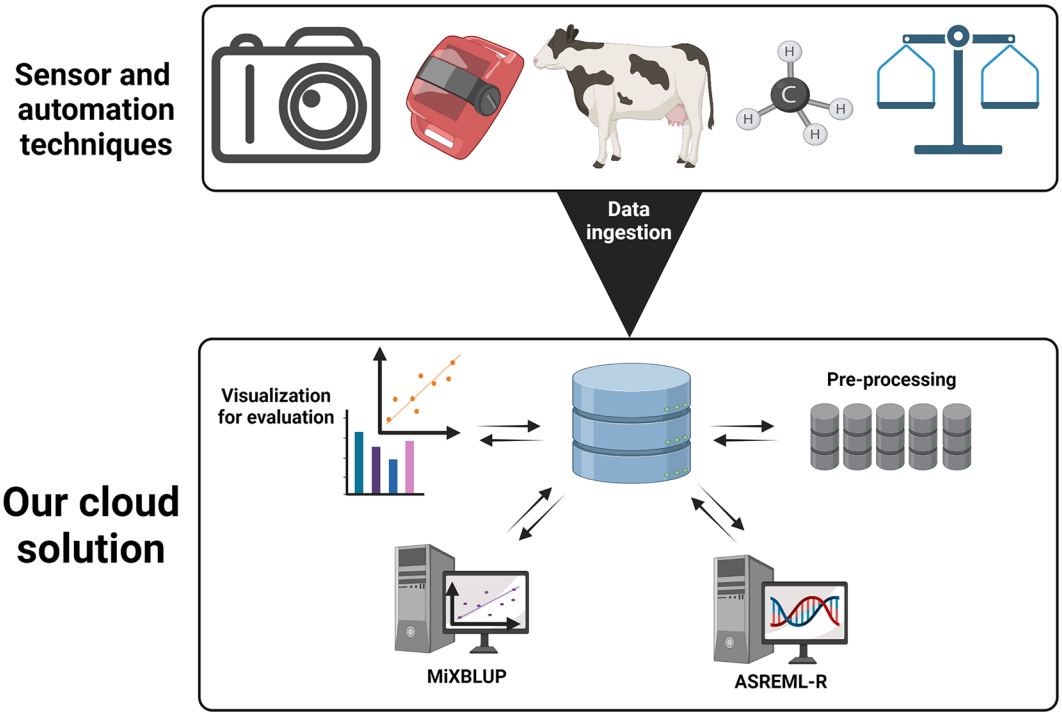

Figure 1. Schematic overview of the types of data, data ingestion, and cloud infrastructure. Within the cloud we have shown the different steps in clock-wise fashion.

the variance of the deviations to the average milk yield per day; and fourth, transforming this variance with a natural logarithm (lnvar), which upon visual inspection made the trait normally distributed. Similar to Poppe et al. (2020), this lnvar was considered the resilience indicator of interest, where a cow with a high value for lnvar is assumed to be less resilient than a cow with a low value for lnvar. In addition, other parameters useful for the genetic statistical model were calculated: herd-year-season and lactation length classes. The ETL procedure resulted in 352,871 cows with one lnvar for each of 3 lactations for each cow. This ETL script is available at https://github.com/dirkjanschokker/cloudsolutionAnimalbreeding.git.

Estimation of Genetic Parameters

A trivariate model was fitted in which lnvar in lactations 1, 2, and 3 was treated as a separate trait. This model included the following fixed effects: herd-year-season, calving age in months, and lactation length (remaining number of days after removing the first and last 10 DIM) in 7 classes, each containing a range of 40 d (50 to 90 d, 91 to 130 d, and so on). The following assumptions were made about the additive genetic effects in the multivariate models:

$$\begin{bmatrix}\mathbf{a_1}\\ \mathbf{a_2}\\ \mathbf{a_3}\end{bmatrix} \sim N\left[\begin{pmatrix}0\\0\\0\end{pmatrix}, \mathbf{A}\otimes\begin{pmatrix}\sigma^2_{a_1} & \sigma_{a_1a_2} & \sigma_{a_1a_3}\\ \sigma_{a_1a_2} & \sigma^2_{a_2} & \sigma_{a_2a_3}\\ \sigma_{a_1a_3} & \sigma_{a_2a_3} & \sigma^2_{a_3}\end{pmatrix}\right],$$

where $\mathbf{A}$ is the additive genetic relationship matrix, $\mathbf{a}_i$ represents the vector with additive genetic effects for trait i, $\sigma^2_{a_i}$ represents the additive genetic variance of trait i, and $\sigma_{a_ia_j}$ represents the genetic covariance between traits i and j. The following assumptions were made about the residuals in the multivariate models:

$$\begin{bmatrix}\mathbf{e_1}\\ \mathbf{e_2}\\ \mathbf{e_3}\end{bmatrix} \sim N\left[\begin{pmatrix}0\\0\\0\end{pmatrix}, \mathbf{I}\otimes\begin{pmatrix}\sigma^2_{e_1} & \sigma_{e_1e_2} & \sigma_{e_1e_3}\\ \sigma_{e_1e_2} & \sigma^2_{e_2} & \sigma_{e_2e_3}\\ \sigma_{e_1e_3} & \sigma_{e_2e_3} & \sigma^2_{e_3}\end{pmatrix}\right],$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{e}_i$ represents the vector with residuals for trait i, $\sigma^2_{e_i}$ is the residual variance of trait i, and $\sigma_{e_ie_j}$ represents the residual covariance between traits i and j. Estimation of variance components is notoriously computationally intensive; therefore, we split the full data set in 5 subsets. This split into 5

subsets was done at the herd level, and each subset contained approximately 20% of the data. Furthermore, we pruned the pedigree file according to a custom-made script using the R package *dplyr* (v1.0.7). This pruning entailed, first, selecting the cows of each subset and, second, including 5 generations of ancestors of the selected cows.

Breeding Value Estimation and Approximation of Reliabilities

The model we used in MiXBLUP was equivalent to the model used in ASReml-R. However, in contrast to the previous step, in which the data were divided into 5 subsets to estimate genetic parameters, the entire data set was used in this step to estimate the breeding values and to approximate reliabilities. Variance components were taken as the average from the subsets. MiXBLUP was used to calculate the reliabilities using the method of Tier and Meyer (2004). The input files for MiXBLUP are also available at https://github.com/dirkjanschokker/cloudsolutionAnimalbreeding.git.

Validation Study

To assess the value of our cloud infrastructure to go from raw (sensor) data to EBV depends on the validation accuracy of these EBV. Therefore, we implemented a 5-fold leave-one-out cross-validation (**LOOCV**) by removing the cows of the earlier used 5 subsets, and we ran the EBV calculations separately 5 times. For each bull, we calculated the mean daughter yield deviations (**DYD**) and the number of DYD per bull in each of the 5 subsets. To have informative data per bull, we only retained bulls that had more than 25 daughters with DYD in each subset, and bulls with an EBV reliability > 0.25 in each subset. Then, each sire had 5 DYD and 5 breeding values estimated. For validation, we averaged the DYD of the 5 subsets and retained the unique bulls over the 5 subsets (n = 173).

RESULTS

Preprocessing and Estimation of Trait Variance and Covariance Components

As a starting point, we used the same data as Poppe et al. (2020). This data set consisted of 1,782,373,113 milk-yield records from 1,120,550 cows. We generated a Python script to preprocess the data by following the ETL procedure. This resulted in 352,871 cows with 1, 2, or 3 lactations with a value for lnvar. The total time to load and preprocess the raw data into a resilience trait per lactation per cow took approximately 6.5 h with our setup. After preprocessing, we fitted linear mixed models with a limited maximum probability for 3 parities at once for each of the 5 subsets. These subsets consisted of 69,841, 71,843, 69,858, 70,947, and 70,382 cows, respectively. The total computation time to estimate the genetic parameters (variance and covariance) for all 5 subsets together was approximately 16 h. Parity 1 showed the largest additive genetic variance (0.038), error variance (0.139), and heritability (0.3214) of lnvar (Table 1). Strong genetic correlations were observed between different parities for 'lnvar' (Table 2). A genetic correlation of 0.98 was observed between parities 2 and 3. The second highest correlation of 0.93 was between parity 1 and parity 2, whereas the lowest correlation of 0.89 was between parities 1 and 3.

Table 1. Variance components—additive genetic variance $\left(\sigma_a^2\right)$, error variance $\left(\sigma_e^2\right)$, and heritabilities—from the univariate analysis of the trait lnvar (SE in parentheses) per parity[1]

Trait	Parity	σ_a^2	σ_e^2	h^2
lnvar[2]	1	0.038 (0.003)	0.139 (0.003)	0.214 (0.033)
	2	0.027 (0.002)	0.131 (0.002)	0.172 (0.035)
	3	0.027 (0.003)	0.123 (0.003)	0.182 (0.012)

[1]Estimates are weighted means of trivariate analyses of 5 subsets of the data, with the empirical SE in parentheses.
[2]Natural log-transformed variance of deviations from a lactation curve.

Prediction and Validation of Breeding Values

Before analyzing the breeding values, we performed a filtering step that selected only those reliabilities >0.25. First, we calculated the mean EBV per parity per birth year for both bulls and cows (Figure 2). We observed a genetic trend from 1995 to 2013 for bulls and from 1995 to 2017 for cows.

To evaluate the performance of the model used on our data set, we implemented a 5-fold LOOCV. An informative data set consisting of 173 bulls was used to measure the performance of the model by plotting the predictions and the observed data (Figure 3), where we observed an R^2 value of 0.592 for the relationship across the LOOCV.

Table 2. Genetic correlations[1] between parities 1, 2, and 3 for the trait lnvar

Comparison	lnvar[2]
Parities 1 and 2	0.93 (0.03)
Parities 1 and 3	0.89 (0.05)
Parities 2 and 3	0.98 (0.02)

[1]Genetic correlations are weighted means of trivariate analyses of 5 subsets of the data, with the empirical SE in parentheses.
[2]Natural log-transformed variance of deviations from a lactation curve.

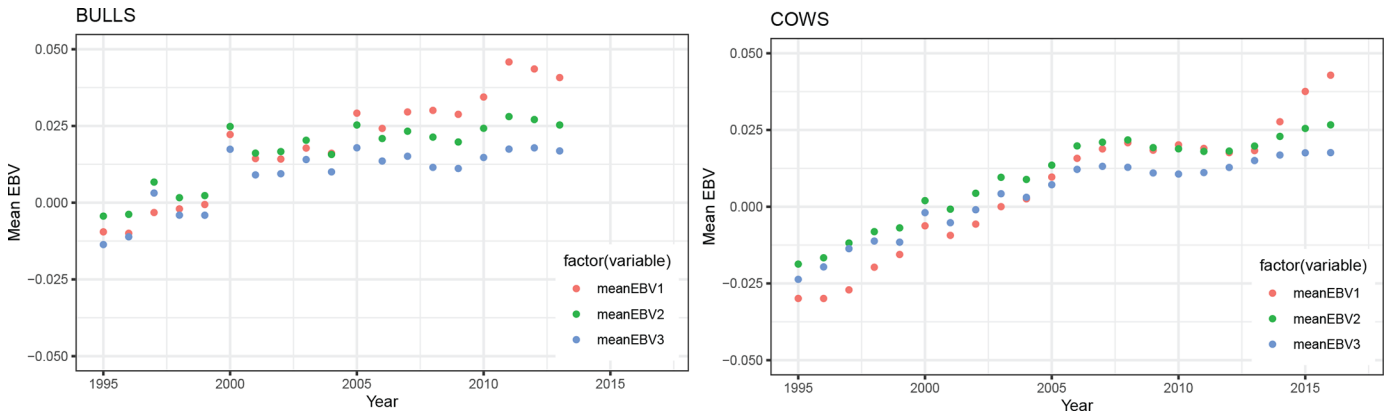

Figure 2. Genetic trend of the mean EBV per year. The x-axis depicts the animal's birth year, whereas the y-axis denotes the mean EBV. Colors indicate the parity, where red is parity 1, green is parity 2, and blue is parity 3. The left panel shows the genetic trend for bulls with an EBV reliability >0.25 and when >500 bulls were present per year. The right panel shows the genetic trend for cows with an EBV reliability >0.25 and when >15,000 cows were present per year.

DISCUSSION

The increasing use of sensor and automation data in the livestock sector provides large volumes of data. Because of the nature of these data, in terms of volume, velocity, and veracity, an alternative infrastructure was proposed and used for storage of such data: a data lake (Schokker et al., 2020). In previous work, we showed part of a working cloud infrastructure in which sensor data from dairy farms, pedigree data, and cow characteristics were incorporated, and subsequent genetic analyses could be performed (Schokker et al., 2022). Here, we go into more detail and discuss the process from choosing the cloud platform to building the infrastructure and demonstrating the functioning of such an infrastructure by applying all the steps necessary to get from raw data to the validation of breeding values using real data. Selection and justification of the computational resources are not trivial, and in our process choices were made to develop a functioning prototype. At the time of designing and building our cloud solution, the main 2 reasons for choosing Microsoft Azure were (1) our already available in-house experience with this platform and (2) the support of the in-house information technology services offered by Wageningen University and Research. In hindsight, we acknowledge that our developed cloud solution could be cost-inefficient at performing certain tasks, such as preprocessing. Nevertheless, we also chose certain tools and specifications (i.e., BLOB storage and GPU) to be ready for other types of sensor and automation data, such as videos or images. BLOB storage has the advantage of storing any type of text or binary data, and therefore is considered to be very useful in housing unstructured data (Bao et al., 2016). The reason for including GPU in our cloud solution was primarily for anticipated future implementation of image or video data that can be used to assess new phenotypes for animal breeding. During development of the current cloud solution, we optimized the cloud infrastructure and worked with real (sensor) data. This was done to demonstrate to animal breeders the potential usefulness and value of our cloud solution in retrieving and validating EBV for traits based on sensor data.

Preparation of the Infrastructure and Implementation of a Novel, Sensor-Based Trait

Defining a novel trait for breeding can be laborious; it encompasses several steps from raw (sensor) data to validating the trait of interest. These steps include data curation, data selection, trait definition, variance component estimation, genetic evaluation, and validation of the EBV. By combining all these steps and parameter settings in one (cloud) environment by using digital (Jupyter) notebooks, we increased flexibility. By "flexibility," we mean the ability to switch easily from running routine genetic evaluations to running new genetic evaluations for new traits based on new sources of information: sensor and automation data. An important step in designing such new evaluations is developing appropriate statistical models. In our study, we used a trait for which the statistical model already had been developed (Poppe et al., 2020, 2021). By implementing an existing statistical model, we could experience and demonstrate the benefits of a cloud solution in the generation of EBV. Benefits of this cloud solution include (1) the possibility for all collaborators (e.g., animal breeders) to upload their raw data in the cloud platform (data lake storage), and (2) the scalability and flexibility of computing power for preprocessing data or running statistical analyses. An example of the scalabil-

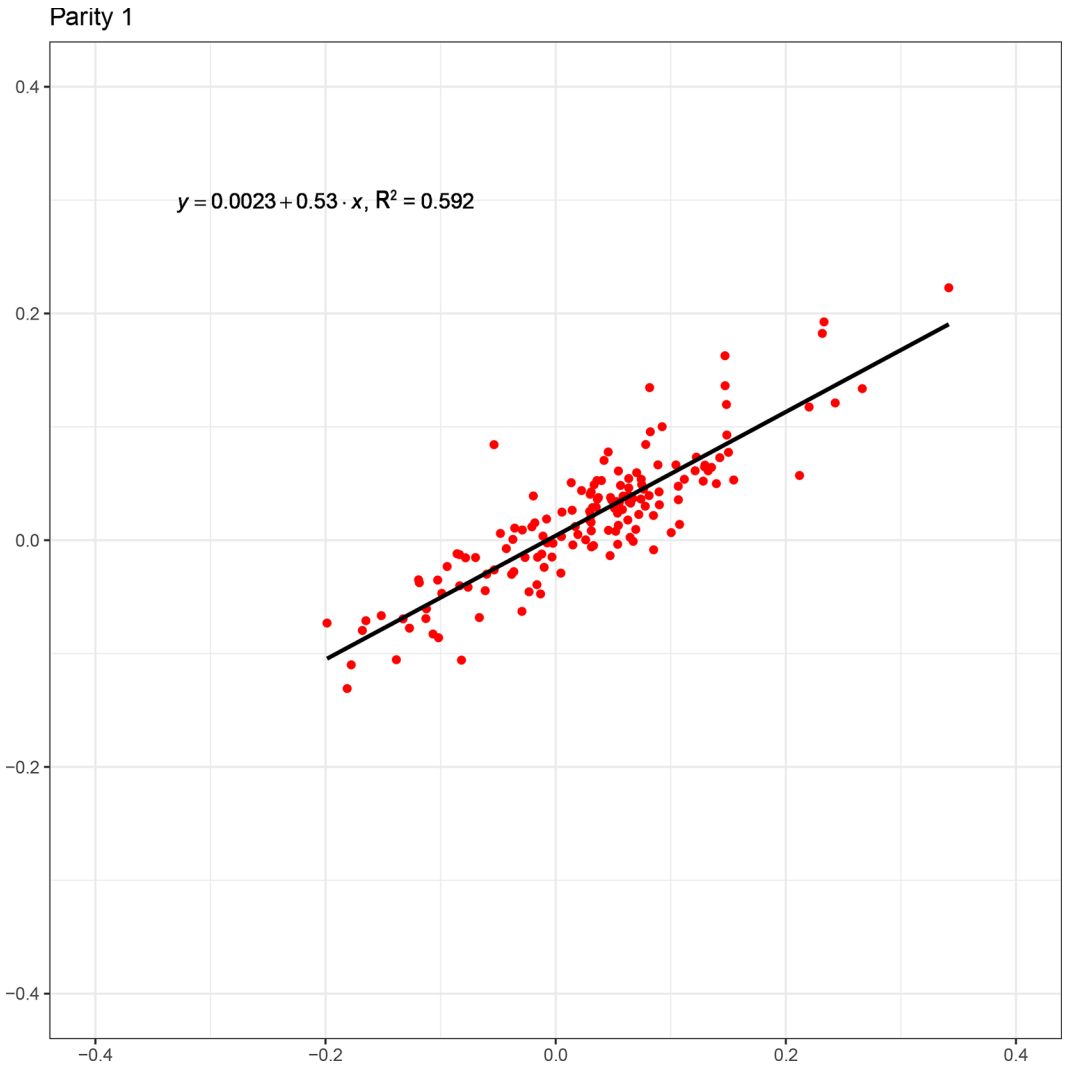

Figure 3. Scatterplots of mean estimated breeding values and mean daughter yield deviations. The x-axis depicts the mean estimated breeding values, whereas the y-axis denotes the mean daughter yield deviations for bulls (n = 173). The resulting R^2 value is 0.592.

ity is the option to run data preprocessing on several clusters, where small but relevant changes in the script can change the outcome of the trait of interest. An example of flexibility includes the option to specify the number of parities to include, the particular genotype on which to focus, or the time period to use (e.g., 1990 through 2000 or 2000 through 2020). Similarly, running downstream analyses simultaneously, such as deriving breeding value estimations with different statistical models, calculating the approximations of reliabilities, and determining yield deviations, are forms of flexibility offered by our cloud solution. To validate whether our developed cloud solution produces EBV comparable to those developed using the more traditional approach, we compared the EBV resulting from our cloud solution with that of prior work, as described next.

Benchmarking of the Results Generated by Our Cloud Solution

Here we discuss the performance of our cloud solution with respect to the statistical models used to generate and validate the EBV with regard to the results of the statistical models. Despite the fact that previous research showed greater genetic improvement of udder health based on polynomial quantile regression when selected on the resilience trait lnvar (Poppe et al., 2020), we chose to include a moving average in our regression model because the moving average showed the highest heritabilities for health, longevity, and fertility traits (Poppe et al., 2020). The additive genetic variance observed in our study was 0.038; the error variance was 0.139. These values are less than the

0.062 (for additive genetic variance) and the 0.192 (for the error variance) reported by Poppe et al. (2020). These differences in values in the variance components could be a result of the optimized scripts used in our study, in which we developed a more detailed equation to calculate the daily milk production per cow by excluding more outliers. The resulting heritability of lnvar of 0.214 in our study was comparable to the 0.244 reported previously by Poppe et al. (2020). Also, the resulting genetic correlations in our study were comparable to those reported by Poppe et al. (2020). These comparable results provide strong support that the implementation of the statistical models in our cloud solution was performed correctly. The validation of novel traits by implementing sensor and automation data has already been performed for traits such as milk yield, weight, mid-infrared spectroscopy (Shetty et al., 2017), ruminating time (Leso et al., 2021), and dry matter intake (Lahart et al., 2019, Martin et al., 2021). The r^2 value observed in our study for lnvar was 0.59, which is in line with the range of values (0.42 to 0.82) for new traits reported in the aforementioned studies (Shetty et al., 2017, Lahart et al., 2019, Leso et al., 2021, Martin et al., 2021).

CONCLUSIONS

We implemented an approach to processing sensor and automation data for defining (novel) traits for animal breeding in which all the required steps to go from raw (sensor or automation) data to EBV are implemented in the cloud. Moreover, our described approach to the exploration and preprocess handling of large volumes of data can be applied to other data and novel traits by modifying our scripts.

ACKNOWLEDGMENTS

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. Because of confidentiality, the data used in this study are not available. The various scripts to load and preprocess the data as well as the input files of MiXBLUP are available and accessible at https://github.com/dirkjanschokker/cloudsolutionAnimalbreeding.git. The authors have not stated any conflicts of interest.

REFERENCES

Bao, J., R. Li, X. Yi, and Y. Zheng. 2016. Managing massive trajectories on the cloud. Page 41 in Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Association for Computing Machinery.

Butler, D. G., B. R. Cullis, A. R. Gilmour, B. G. Gogel, and R. Thompson. 2017. ASReml-R Reference Manual Version 4. VSN International Ltd.

Gengler, N. 2019. Symposium review: Challenges and opportunities for evaluating and using the genetic potential of dairy cattle in the new era of sensor data from automation. J. Dairy Sci. 102:5756–5763. https://doi.org/10.3168/jds.2018-15711.

Lahart, B., S. McParland, E. Kennedy, T. M. Boland, T. Condon, M. Williams, N. Galvin, B. McCarthy, and F. Buckley. 2019. Predicting the dry matter intake of grazing dairy cows using infrared reflectance spectroscopy analysis. J. Dairy Sci. 102:8907–8918. https://doi.org/10.3168/jds.2019-16363.

Leso, L., V. Becciolini, G. Rossi, S. Camiciottoli, and M. Barbari. 2021. Validation of a commercial collar-based sensor for monitoring eating and ruminating behaviour of dairy cows. Animals (Basel) 11:2852. https://doi.org/10.3390/ani11102852.

Martin, M. J., J. R. R. Dorea, M. R. Borchers, R. L. Wallace, S. J. Bertics, S. K. DeNise, K. A. Weigel, and H. M. White. 2021. Comparison of methods to predict feed intake and residual feed intake using behavioral and metabolite data in addition to classical performance variables. J. Dairy Sci. 104:8765–8782. https://doi.org/10.3168/jds.2020-20051.

Matthews, S. G., A. L. Miller, J. Clapp, T. Plotz, and I. Kyriazakis. 2016. Early detection of health and welfare compromises through automated detection of behavioural changes in pigs. Vet. J. 217:43–51. https://doi.org/10.1016/j.tvjl.2016.09.005.

Ouweltjes, W., M. Spoelstra, B. Ducro, Y. de Haas, and C. Kamphuis. 2021. A data-driven prediction of lifetime resilience of dairy cows using commercial sensor data collected during first lactation. J. Dairy Sci. 104:11759–11769. https://doi.org/10.3168/jds.2021-20413.

Poppe, M., G. Bonekamp, M. L. van Pelt, and H. A. Mulder. 2021. Genetic analysis of resilience indicators based on milk yield records in different lactations and at different lactation stages. J. Dairy Sci. 104:1967–1981. https://doi.org/10.3168/jds.2020-19245.

Poppe, M., R. F. Veerkamp, M. L. van Pelt, and H. A. Mulder. 2020. Exploration of variance, autocorrelation, and skewness of deviations from lactation curves as resilience indicators for breeding. J. Dairy Sci. 103:1667–1684. https://doi.org/10.3168/jds.2019-17290.

Schokker, D., I. N. Athanasiadis, M. Poppe, J. ten Napel, C. Kamphius, and R. F. Veerkamp. 2022. From raw sensor and automated data to genetic evaluation and validation in the cloud. World Congress on Genetics Applied to Livestock Production, Rotterdam, the Netherlands. (Abstr.)

Schokker, D., I. N. Athanasiadis, B. Visser, R. F. Veerkamp, and C. Kamphuis. 2020. Storing, combining and analysing turkey experimental data in the Big Data era. Animal 14:2397–2403. https://doi.org/10.1017/S175173112000155X.

Shetty, N., P. Lovendahl, M. S. Lund, and A. J. Buitenhuis. 2017. Prediction and validation of residual feed intake and dry matter intake in Danish lactating dairy cows using mid-infrared spectroscopy of milk. J. Dairy Sci. 100:253–264. https://doi.org/10.3168/jds.2016-11609.

Smith, K., A. Martinez, R. Craddolph, H. Erickson, D. Andresen, and S. Warren. 2006. An integrated cattle health monitoring system. Pages 4659–4662 in Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol. 1. 2007/10/20 ed. IEEE Engineering in Medicine and Biology Society.

Tier, B., and K. Meyer. 2004. Approximating prediction error covariances among additive genetic effects within animals in multiple-trait and random regression models. J. Anim. Breed. Genet. 121:77–89. https://doi.org/10.1111/j.1439-0388.2003.00444.x.

Zaharia, M., M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. 2010. Spark: Cluster computing with working sets. HotCloud 10:95.