A Low-cost Infrastructure for Massive Storage of Phenotype Data for Dairy Cattle Genetic Improvement Programs

Wagner Arbex[*,†], Caio dos Santos Borsato de Carvalho[*,‡], Katia Santos[*],
Vinicius Campista Brum[§] and Marcos Vinícius Barbosa da Silva[*]
[*]Brazilian Agricultural Research Corporation (Embrapa)
Juiz de Fora, MG, Brazil
[†]Correspondent author – wagner.arbex@embrapa.br
[‡]National Council for Scientific and Technological Development (CNPq) grant
[§]Federal University of Juiz de Fora (UFJF)

***Abstract*—The activities and progress of genetic breeding were always related to computing – or rather, the availability of appropriate computational resources to the behavior of genetic improvement, because, either by cross breeding or selection, the data sets involved in genetic improvement research and development activities care both in quantity and in quality. The demand for computing resources in these programs is extensive and intense, because genetic-genomic evaluation requires greater availability and accuracy of phenotype data, given the use of assessment models increasingly sophisticated and, moreover, are added to these aspects the need to interpret and understand the phenotype databases from a logical and efficient structure for data storage and information retrieval. This paper presents a low-cost solution to implement a storage system for huge data mass.**

***Index Terms*—Scientific computing environment, storage system, utility computing, phenotype data, genetic improvement**

I. Introduction

Historically, the works and progress of genetic improvement were always related to computing – or rather, the availability of appropriated computational resources to the behavior of genetic improvement [1]. This fact is due to the nature of research activities, because either by cross breeding or selection, the data sets involved in genetic improvement research and development activities care both in quantity and in quality of these.

Addition to the traditional approach of the breeding programs, currently works with genomic selection are being developed, complementing the work already developed and allowing investigation of the genetic potential of individuals before they can express their phenotypic characteristics.

Over again, the demand for computational resources is extensive and intense, because genetic-genomic evaluations require greater availability and accuracy of phenotype data, given the use of assessment models increasingly sophisticated.

Add up to these aspects the need to interpret and understand the phenotype databases from a logical and effective structure for data storage and information retrieval.

Initiatives for data storage in genetic breeding programs are not unknown. Among others, the National Dairy Cattle Research Center (Embrapa Dairy Cattle) of the Brazilian Agricultural Research Corporation (Embrapa) has worked successfully. For example, computing systems were developed for the creation and implementation of the National Animal Science Archive of Dairy Cattle (Arquivo Zootécnico Nacional de Gado de Leite – AZN-GL) estabilished by the Ministry of Agriculture of Brazil through Ordinance #33 on February 10, 1987 [2] .

Now, this work shows a free-software based storage solution to phenotype data warehousing currently being implemented to the Embrapa Dairy Cattle's genetic improvement programs.

II. Data science approach

The computer science and the information systems, utilized as tools by science, changed the agriculture and livestock fields the same way as they did with many other fields of science.

Therefore, considering the exponentially increasing amount and the high complexity of scientific data that are being generated and which need to be efficiently processed, new computational resources are required for effective treatment of this entire data volume, so they can be transformed into knowledge, enabling the adhibition and allowing or enhacing technological advances in order to promote the upgrading of the productive sectors.

The use of mathematical and computational models as a research tool not only makes the interpretation of the handled easily content, but also of complex data sets currently generated, whose characteristics include, inter alia [3]:

- large data volume, which data sets of terabytes magnitude are becoming usual;

- high dimensionality, when working with hundreds or thousands of attributes to be studied;
- heterogeneity, since unlike traditional methods of analysis, computational models are suitable for different data types, discrete and uncategorized;
- multiple physical location of data sets, since it is common these databases are distributed and/or dispersed in different repositories.

III. Motivation and earlier programs

A pioneering initiative to capture phenotype primary data was the development of PROLEITE – a computer software for which three versions were developed [4], [5], [6] – and its use was discontinued in the 90s, partially motivated by the fact that its biggest users, the breeders' associations, began to develop their own applications and/or get them from third-party. These new software could meet their needs and also fulfill the requirement to meet the demand for inclusion of phenotype data in AZN-GL.

The AZN-GL has increased its data volume by 40% per year (Figure 1[1]), at least, and has established itself as a source of data from numerous studies – coinciding with the first studies in Brazil using genetic evaluation software which have implemented the Animal Model.

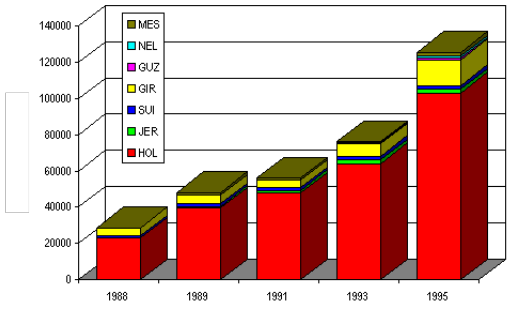

Fig. 1. The AZN-GL growth in the early years of the bovine datasets of the breeds Nelore (NEL), Guzera (GUZ), Gyr (GIR), Brown Suiss (SUI), Jersey (JER), Holstein (HOL) and crossbreed (MES) [7].

In the second half of the 90s, beginning a new venture, the work on the AZN-GL and the PROLEITE became joint projects between the Embrapa Dairy Cattle and the Embrapa Agricultural Informatics [8], [9]. Subsequently, the initiative of the PROLEITE, with new dimensions, was resumed by these research centers [10].

However, the main issue for the maintenance and existence of the AZN-GL was not getting the primary data, but the strategies and methods for the structure and database processing. These strategies and methods should allow the database usage for genetic evaluations implementation – among other analyzes and other research projects investigations.

[1]Nelore, Guzera and Gyr there are Zebu subspecies.

Thus, as another action strategy for the deployment of AZN-GL and genetic evaluations running, a computer system, named Crivo [11], with automatic or semiautomatic methods was developed. These methods were defined in protocols for inclusion, verification, modification, query and retrieval of data.

The methods and strategies deployed in the Crivo proved to be fundamental for all data manipulation of the AZN-GL and, moreover, allowed that their databases were reliable, undamaged and available to be used in several works of the animal breeding research team of Embrapa Dairy Cattle.

Many studies could only be made through the data provided by the AZN-GL. Some of these works are the earliest texts in this field in Brazil, and showed up, e.g., different evaluation models, including the first evaluations with the Animal Model in the country, new studies of reproductive and productive efficiency, the estimation of new parameters for breeds and specific conditions into Brazil, or even the first adjustment factors for lactations also for milk production specific characteristics in the country.

Genetic evaluations using the Animal Model are examples of the importance and necessity of adequate computing resources to process phenotype data. This model, though known since the 60s, became to be used in the late 80s due to the lack of sufficient resources. In Brazil, the Animal Model became to be used in the following decade.

Particularly, new computational resources to perform genetic evaluations using the new Animal Model were required to meet their data preparing needs, and so were added to Crivo new methods that allowed the generation of these files using the new evaluation models template.

It should be emphasized that, at that time, the available computational resources in Brazil were poor – when compared to those used in United States, Canada and European countries that allowed the data processing with some "facility".

In order to, at that time, was necessary to implement non-trivial algorithms that could allow the manipulation of data masses using low computing resources. For example, algorithms for partial and topological sort had been implemented to reduce the memory use, disk access, and thus the runtime.

Algorithms with such implementations were unknown for this purpose – in the case of large data sets – and for small sized hardware. As an example, the development and implementation of such algorithms allowed the work completion conducted by Silva [12] and Houri Neto [13].

The computational activities for Houri Neto's [13] thesis were particularly challenging once datasets with a huge number of entries, up to that time, had not been collected for a single breed in Brazil. In their origin – outside of Brazil – these data sets were hosted in large sized computer equipment, named mainframes, and manipulated by

database managers systems (DBMS) commercially established in the market, which was not true in Brazil.

For these reasons, in the past or at the present, the implementation of efficient computational methods that are capable of dealing with this massive amount of data and allow analyze them reliably is characterized as one of the challenges in breeding genetics programs.

IV. Infrastructure for massive storage datasets

Cloud computing is a new business model of IT, but also a new way of computing resources organization where can be made available software, platform, infrastructure, storage area, inter alia. Moreover, it can also provide the desktop computer itself to the end-user. That is, create and provide the desktop resources to users and clients from easy access procedures.

Both technical and operational aspects of cloud computing are based on utility computing, which in turn is structured by a subset of the distributed systems and operating systems concepts and tools, e.g., distribution transparency and virtualization [14], [15].

Storage systems are excellent solutions for scientific computing environments, which usually, the nature of the activity hinders to predict the data volume to be treated. Furthermore, storage systems are expensive solutions and might require changes in the computing environment.

The solution presented was developed using four CPUs with Linux operating system, creating a disk array with capacity of over 4 TB of raw space. To implement the disk array was used MHDDFS [16] and NFS [17] applications, available for Linux and other operating systems. The NFS layer provides the communication among the hard disks over the network and the MHDDFS layer allows the union of several mount points in a single space, assembling the disk array.

Integrated to the storage system was implemented a backup system that copies the whole storage contents, four times a day, in external hard drives with network access. The backup system also had 4 TB of workspace, and it was implemented with native Linux tools – such as CRON and RSYNC – it also has its operation without the intervention or even the user's knowledge.

This storage system was implemented with free software and low-cost hardware, but, despite this, the logical organization and the working structure of the cloud computing allows scalability and security in storage to scientific computing environments and computing on huge databases in general.

Currently, this storage system has been upgraded. It has been improved in the backup mechanism and storage capacity. The storage system area was increased in about 65% and the backup area hard disks was quadruplicate. Respectively, were about 6.5 TB and 16 TB.

Another feature that should be modified is the access way for users, long as new interface and access mechanism will be developed.

V. Conclusions

The efficient and appropriate phenotypic data storage always were problems for genetic improvement programs. Research groups around the world need and are looking for a solution for this issue.

The solution is not restricted to a computational infrastructure for storage resources but begins from this. From an infrastructure point of view this paper presents a low-cost solution using free software and/or open source resources that can be implemented even in obsolete or unsophisticated hardware if necessary. In addition, as a concept of cloud computing and distributed systems, provides storage as a service (StaaS), implementing transparency of access, location, migration, replication, and concurrence.

Possibly, its biggest disadvantage is not to assure a great throughput for data access. Because it does not use an independent or own network structure as many commercial storage systems. However, there are network architectures that can reduce this problem.

The atributes and features offered by this solution can fully meet the needs of massive data storage in genetic breeding programs. Among these, stand out from advantages the ease of implementation, configuration and scalability.

Acknowledgments

The authors thanks to reviewers who gave useful comments, and to funding agencies and corporate sponsors that provided financial support, which will be nominated in the camera-ready version.

References

[1] H. Grosu, L. Schaeffer, P. A. Oltenacu, D. Norman, R. Powell, V. Kremer, G. Banos, R. Mrode, J. Carvalheira, J. Jamrozik, C. Draganescu, and S. Lungu, *History of genetic evaluation methods in dairy cattle*, Jan. 2013.

[2] BRASIL., "Institui o arquivo zootécnico nacional," Diário Oficial, Seção I. Portaria n. 32, de 10 de fevereiro de 1987, Brasília, p. 2126, Fev. 1987.

[3] P. Tan, M. Steinbach, and V. Kumar, *Introdução ao data mining: mineração de dados*. Rio de Janeiro: Ciência Moderna, 2012.

[4] W. Arbex and M. L. Martinez, "PROLEITE/PROLEI-PC 1.0," 2 disquetes, Juiz de Fora, 1989, c e Pascal. Ambiente DOS.

[5] W. Arbex, M. L. Martinez, and M. V. L. dos Santos, "PROLEITE/PROLEI-PC 2.0," 2 disquetes, Juiz de Fora, 1990, c e Pascal. Ambiente DOS.

[6] W. Arbex, M. V. L. dos Santos, and M. L. Martinez, "PROLEITE/PROLEI-PC 3.0," 2 disquetes, Juiz de Fora, 1994, c e Pascal. Ambiente DOS.

[7] S. R. de Medeiros Oliveira, W. Arbex, C. N. Costa, M. A. Arbex, and W. C. P. de Magalhães Júnior, "Automação dos processos de gestão do Arquivo Zootécnico Nacional de Gado de Leite (AZN-GL)," in *Anais...*, Agrosoft 99 – II Congresso da Sociedade Brasileira de Informática Aplicada à Agropecuária e Agroindústria. Campinas: Sociedade Brasileira de Informática Aplicada à Agropecuária e Agroindústria, 1999. [Online]. Available: http://www.agrosoft.org.br/trabalhos/ag99/artigo07.htm

[8] S. R. de Medeiros Oliveira, C. da Costa Carrer, W. Arbex, and J. Valente, "Sistema de análise e acompanhamento de produção de rebanhos leiteiros," in *Anais...*, XXXIV Reunião da Sociedade Brasileira de Zootecnia. Juiz de Fora: Sociedade Brasileira de Zootecnia, Ago. 1997, pp. 338–339.

[9] S. R. de Medeiros Oliveira, M. Pedroso Júnior, W. Arbex, and C. da Costa Carrer, "PROLEITE: sistema de análise e acompanhamento de rebanhos leiteiros," in *Anais...*, Agrosoft 97 – I Congresso da Sociedade Brasileira de Informática Aplicada à Agropecuária e Agroindústria. Belo Horizonte: Sociedade Brasileira de Informática Aplicada à Agropecuária e Agroindústria, 1997, pp. 141–147. [Online]. Available: http://www.agrosoft.org.br/trabalhos/ag97/c3t1100.htm

[10] S. R. de Medeiros Oliveira, C. N. Costa, and C. C. Sabadini, "Sistema para auxiliar o manejo de rebanhos leiteiros e a tomada de decisões," Embrapa Informática Agropecuária, Campinas, Comunicado Técnico 1, Abr. 1999.

[11] W. Arbex and M. L. Martinez, "Crivo," 1 CD, Juiz de Fora, 1996, c. Ambiente Unix-like.

[12] M. V. G. B. da Silva, "Utilização de modelos uni e bivariados no estudo das relações entre eficiência reprodutiva e produção de leite na raças holandesa," Mestrado em Ciências em Zootecnia, Universidade Federal de Minas Gerais, Belo Horizonte, 1995.

[13] M. Houri Neto, "Interação genótipo-ambiente e avaliação genética de reprodutores da raça Holandesa usados no Brasil e nos Estados Unidos da América," Doutorado em Ciência Animal, Universidade Federal de Minas Gerais, Belo Horizonte, 1996.

[14] W. Arbex, R. F. Tagliatti, L. G. de Andrade, M. N. M. Muniz, E. Guedes, and M. V. G. B. da Silva, "Storage as a service and cloud computing for bioinformatics computing environment," in *Anais...*, X-meeting 2010 – VI International Conference of the Brazilian Association for Bioinformatics and Computational Biology. Ouro Preto: Universidade Federal de Ouro Preto - UFO, Out. 2010, p. 103.

[15] W. Arbex, M. V. B. da Silva, M. F. M. Guimarães, R. F. Tagliatti, L. G. de Andrade, M. N. M. Muniz, and L. A. V. de Carvalho, "Storage as a service and utility computing for bioinformatics computing environment: aspects of cloud computing applied to scientific computing," in *Anais...*, IV Encontro Acadêmico em Modelagem Computacional. Petrópolis: Laboratório Nacional de Computação Científica - LNCC, Jan. 2011, trabalho Destaque em Ciência da Computação no IV Encontro Acadêmico em Modelagem Computacional do Laboratório Nacional de Computação Científica.

[16] D. E. Oboukhov, "MHDDFS," 1 file, 2008, MHDDFS – Multiple Hard Disk Distributed File System. [Online]. Available: http://mhddfs.uvw.ru/

[17] S. Shepler, B. Callaghan, D. Robinson, R. Thurlow, Sun Microsystems Inc., C. Beame, Hummingbird Ltd., M. Eisler, D. Noveck, and Network Appliance Inc., "Network File System (NFS) version 4 Protocol," IETF – The Internet Engineering Task Force, Fremont, Tech. Rep. 3530, Dec. 2000. [Online]. Available: http://tools.ietf.org/html/rfc3530