IOSR Journal of Business and Management (IOSR-JBM)
e-ISSN:2278-487X, p-ISSN: 2319-7668. Volume 26, Issue 12. Ser. 12 (December. 2024), PP 22-29
www.iosrjournals.org

Data Warehouse For The State Of Rondônia Milk Production Chain: Construction Of Algorithms For Data Extraction, Transformation And Loading

Kaio Alexandre Da Silva[1]*, Márcio Rodrigues Miranda[1],
Luiz Francisco Machado Pfeifer[2]
[1]*Instituto Federal De Educação, Ciência E Tecnologia De Rondônia (IFRO), Porto Velho, Rondônia, Brazil*
[2]*Empresa Brasileira De Pesquisa Agropecuária (EMBRAPA), Porto Velho, Rondônia, Brazil*

Abstract:
The dairy production chain plays a strategic role in global agribusiness. In Brazil, it plays a prominent role, being one of the most important for the national economy. Brazil has been standing out as one of the world's largest milk producers, being the largest milk producer in South America and the fifth largest producer in the world, according to 2023 production. This study presents the creation of a Data Warehouse to analyze the milk production chain in the State of Rondônia, located in the Brazilian Amazon region. The methodology used was Design Science Research, which focuses on building a technological artifact to solve the problem of falling production. The data was extracted from several official sources from Brazilian agencies. The Extraction, Transformation and Loading process was carried out using the Python programming language, the database used was MySQL and the data visualization was implemented with the Metabase tool. The Data Warehouse allowed the integration and analysis of data on production, value, herd, pasture area and vaccination, provided through dashboards created with Metabase, enabling the dynamic and interactive visualization of regional scenarios. The results contribute to the understanding of the factors that influence milk production in Rondônia and can help in the formulation of public policies and strategies for the sustainable development of the production chain.

Key Word: *Data Warehouse; Dashboards; ETL; Dairy Cows.*

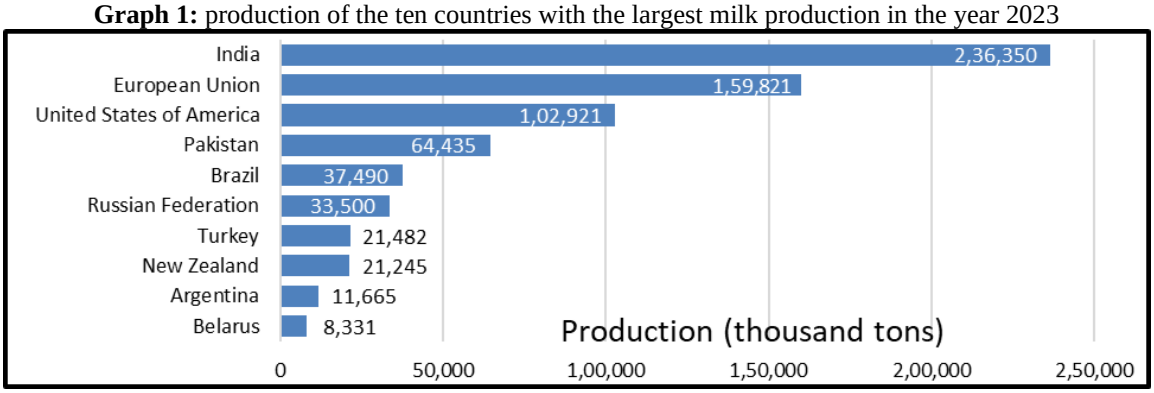
Date of Submission: 08-12-2024 Date of Acceptance: 18-12-2024

I. Introduction

The dairy production chain plays a strategic role in global agribusiness. In Brazil, it plays a prominent role, being one of the most important for the national economy. Brazil has been standing out as one of the world's largest milk producers, being the largest milk producer in South America and the fifth largest producer in the world, according to the 2023 production, as shown in

Graph ***1***, which shows the production of the ten countries with the largest milk production in the year 2023[1].

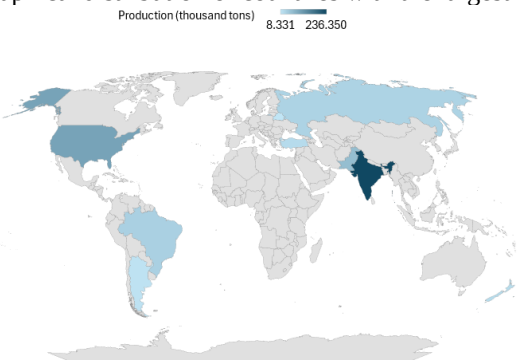
Graph 1: production of the ten countries with the largest milk production in the year 2023



Milk and dairy production are a fundamental pillar of the global economy and social well-being. Dairy products are an irreplaceable source of essential nutrients such as calcium, protein and vitamins, ensuring food and nutritional security, especially in developing countries. India, as the world's largest producer, exemplifies

this impact, with large-scale production boosting its economy and providing livelihoods to millions of farmers and workers. Figure 1 shows the geographical distribution of countries with the largest milk production[1].

Figure 1: geographical distribution of countries with the largest milk production



In Brazil, milk production is intrinsically linked to the country's regional structure, reflecting variations in climate, infrastructure, access to markets and production characteristics. Brazil is divided into 27 states, distributed among 5 regions, as shown in

Figure ***2***.

Figure 2: Brazilian States

Milk production in Brazil is concentrated in states such as Minas Gerais, Paraná and Rio Grande do Sul, while other regions, such as the Northeast, have shown increasing participation in the sector due to incentive policies and technological improvements[2].

Given the economic and social relevance of the dairy production chain, understanding the regional distribution of production and the factors that drive or limit its expansion is essential for planning public policies and sustainable development strategies, especially for states in the Amazon region, which are characterized as difficult to access in Brazil.

This study will present the creation of a Data Warehouse (DW) for the production chain of the state of Rondônia. Rondônia is a state located in the north of Brazil and is one of the states that make up the Brazilian Amazon region. The need to create a DW for the production chain of Rondônia is since in the last 5 years, Rondônia's milk production has fallen by approximately 43%. In 2019, Rondônia was the seventh largest milk producer in the country, producing approximately 1.128.597.000 liters of milk, however, in 2023 it dropped to twelfth place, producing approximately 644.192.000 liters of milk. Table 1 shows the ranking of the 10 main milk producing states in Brazil between 2019 and 2023[2].

Table 1: Shows the ranking of the 10 main milk producing states in Brazil between 2019 and 2023

Production Milk (thousand liters)								
Position	State	2019	Position	UF	2020	Position	UF	2021
1	Minas Gerais	9.447.532	1	Minas Gerais	9.692.389	1	Minas Gerais	9.611.706

2	Paraná	4.349.171	2	Paraná	4.671.014	2	Paraná	4.415.634
3	Rio Grande do Sul	4.302.968	3	Rio Grande do Sul	4.249.805	3	Rio Grande do Sul	4.400.356
4	Goiás	3.164.963	4	Goiás	3.173.510	4 - ↑ 1	Santa Catarina	3.161.993
5	Santa Catarina	3.040.179	5	Santa Catarina	3.137.219	5 - ↓ 1	Goiás	3.121.391
6	São Paulo	1.653.036	6	São Paulo	1.645.653	6	São Paulo	1.570.502
7	Rondônia	1.128.597	7 - ↑ 1	Bahia	1.064.599	7	Bahia	1.202.553
8	Bahia	1.069.019	8 - ↑ 1	Pernambuco	1.036.756	8	Pernambuco	1.137.809
9	Pernambuco	1.055.790	9 - ↓ 2	Rondônia	925.803	9 - ↑ 1	Ceará	960.436
10	Ceará	797.362	10	Ceará	871.529	10 - ↓ 1	Rondônia	741.053

Position	UF	2022	Position	UF	2023
1	Minas Gerais	9.362.690	1	Minas Gerais	9.422.028
2	Paraná	4.459.359	2	Paraná	4.557.624
3	Rio Grande do Sul	4.071.675	3	Rio Grande do Sul	4.114.178
4	Santa Catarina	3.134.943	4	Santa Catarina	3.205.790
5	Goiás	2.995.345	5	Goiás	2.980.911
6	São Paulo	1.524.256	6	São Paulo	1.511.847
7	Bahia	1.276.259	7 - ↑ 1	Pernambuco	1.334.730
8	Pernambuco	1.165.546	8 - ↓ 1	Bahia	1.267.223
9	Ceará	1.067.309	9	Ceará	1.135.748
10	Rondônia	655.790	10 - ↑ 1	Alagoas	703.448

Therefore, this study presents the methodology used to create the DW, presenting the process of extracting, transforming and loading data from the Municipal Livestock Survey carried out by the Brazilian Institute of Geography and Statistics (IBGE), as well as data from the Herd Declaration Campaign issued by the Agrosilvopastoral Health Defense Agency of the State of Rondônia and mapping data of areas occupied by pastures in the State of Rondônia, made available by the Atlas of Brazilian Pastures.

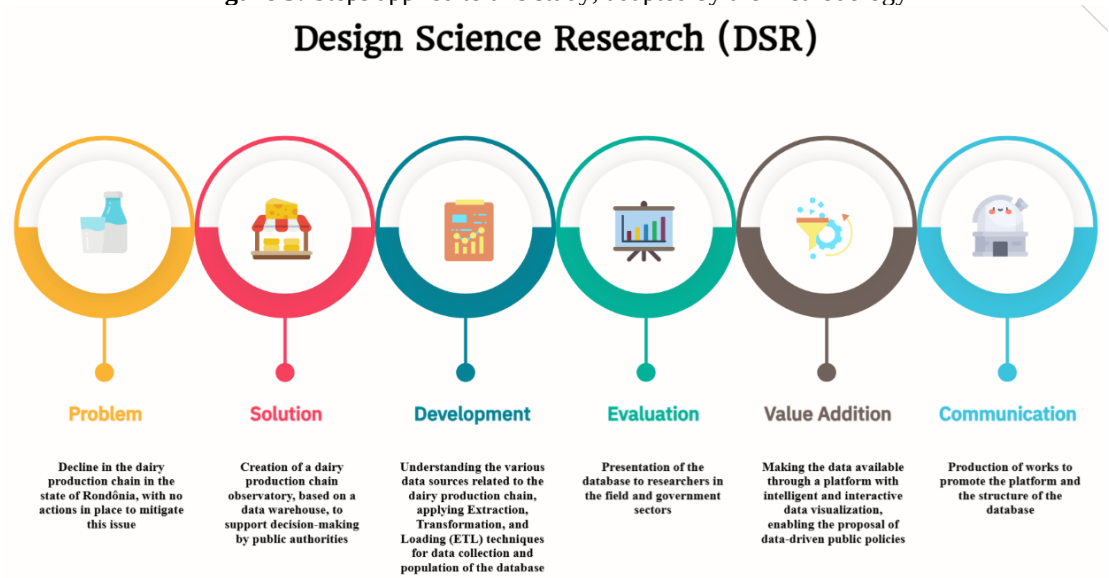
II. Material And Methods

The methodology adopted to create the DW was Design Science Research (DSR). Originating from the area of Design Science, the DSR methodology has as its main objective the creation and evaluation of technological artifacts that seek to solve real and relevant problems, promoting innovations or significant improvements in practical contexts[3].

DSR is a research methodology applied in the development of technological solutions, being widely used in areas such as computer science, engineering and information systems. The methodology is structured in cycles of design, construction, evaluation and refinement, allowing the proposed solution to be continuously improved based on its practical application and the results obtained[4].

Figure ***3*** presents the steps applied to this study, adapted by the methodology.

Figure 3: Steps applied to this study, adapted by the methodology



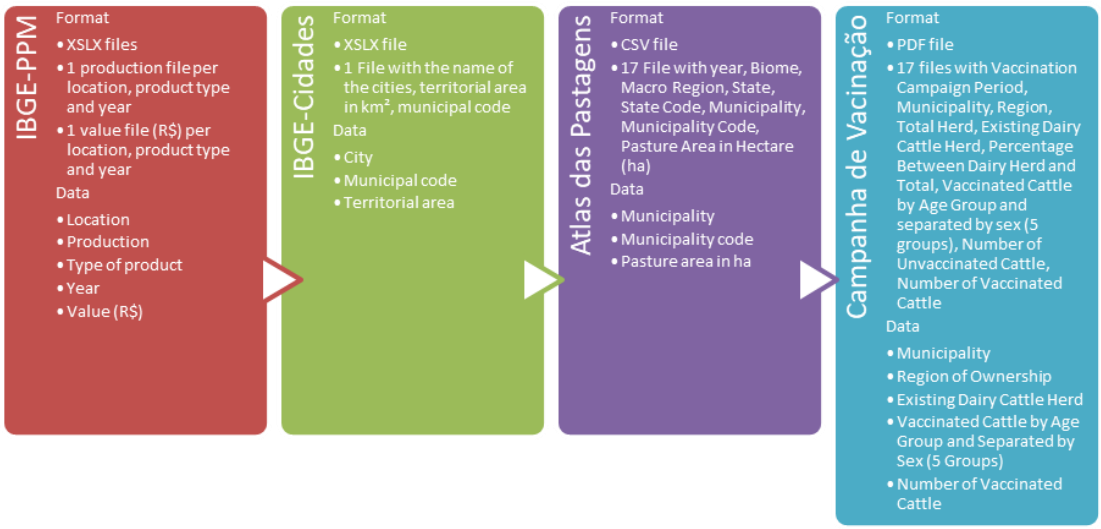
The following tools were used: Python programming language, version 3.10, used to perform the data extraction, transformation and loading process. To do this, it was necessary to use the Pandas and Tabula libraries, so that it would be possible to read documents in PDF format and read files in CSV and XLSX format. MySQL was used as the database, which is a structured database that uses the entity-relationship model. To visualize the data, the Metabase tool was used, which is an open-source tool for querying, visualizing and analyzing data in a simple and accessible way. For the ETL process, the first step was to identify which databases would be necessary for data collection, which were established according to the data to be made available:

- IBGE – PPM[2]: Municipal Livestock Survey is a government database that covers livestock numbers in the municipality on the reference date of the survey, as well as animal production and the value of production during the reference year. The numbers include cattle, pigs, swine breeding stock, poultry, chickens, quails, horses, buffaloes, goats and sheep. Animal production, in turn, includes the production of milk, chicken eggs, quail eggs, honey, raw wool and silkworm cocoons; the quantities of milked cows and shorn sheep; and aquaculture, which includes fish farming, shrimp farming and malacoculture. Held annually, it has a national geographic scope, with results published for Brazil, Major Regions, Federation Units, Mesoregions, Microregions and Municipalities;
- IBGE – Cities[5]: Cities and States of Brazil is an information system about the municipalities and states of Brazil, which has data on various topics and research in the form of tables, graphs and maps, as well as historical information and photographs. It is also possible to compare information between different municipalities and states;
- Atlas of Pastures[6]: It is a platform that provides and organizes databases on the mapping of areas occupied by pastures in Brazil, mapping the quality of these areas, estimates of carbon stocks in pastures in the Cerrado biome, information on the Brazilian cattle herd analyzed based on data from the municipal livestock survey (IBGE-PPM), time series of land use and land cover classification by visual inspection of data from Landsat satellites, assessment of the condition of pastures in field campaigns carried out by laboratory teams;
- Vaccination Campaign[7]: IDARON is a government agency of the State of Rondônia, whose responsibilities include the health control of cattle herds, the Health Inspection of Products and By-products of Animal Origin, and the Inspection and Control of the Transit of Animal Material. Therefore, among its operating systems, measures were established such as monitoring through system declarations, animal vaccination campaigns, registration base and audit of the agricultural production system, veterinary care, active and passive epidemiological surveillance. In this way, making the data available for each vaccination campaign.

After identifying the databases to be searched, the data that would be collected from each database were mapped, as well as the cleaning, standardization and transformation necessary for the effective storage of the data.

*Figure **4*** shows the organization of each database, as well as the file formats provided and the attributes belonging to each database.

Figure 4: shows the organization of each database



III. Discussion

The implementation of the ETL process for the Data Warehouse of the milk production chain in Rondônia allowed the integration and analysis of multiple data sources. The process was structured in three main stages: Extraction, Transformation and Loading, as shown in the flowchart presented in

Figure ***5***.

In the Extraction phase, the main challenge was in the IBGE-PPM and Vaccination Campaign files. Starting with the IBGE-PPM data structure, the data is in XLSX format, separated into internal spreadsheets, with an unorganized structure, as shown in Figure 6 (a, b). The first step was to clean the spreadsheets, deleting header lines 1 to 4, to reduce the complexity of data transfer, and this procedure was performed in both spreadsheets. The next step was to define which partial structure would be used, since the years were being used as attributes and not values, so it was established that a partial CSV file would be generated, which would merge the data from the two spreadsheets, condensing values as shown in Figure 6 (c). In this way, we ensured that we could perform queries in the location and year dimensions. In the Vaccination Campaign files, the data is in PDF format, presenting a consolidated data table. Figure 7 shows the table formatting. However, the files were not standardized between years. One of the reasons for this may have been the use of different conversion engines for each reporting year. Therefore, since it was not possible to clean this file, the strategy adopted was to create an algorithm in Python that receives the file and, through the "Tabula" library, searches for the data.

Figure 5: Flowchart of ETL proccess

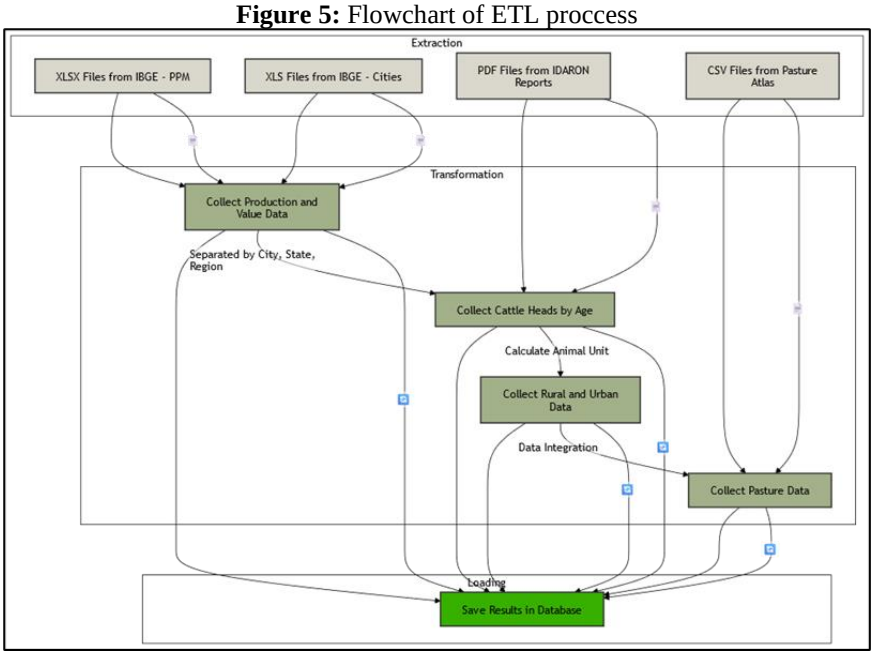
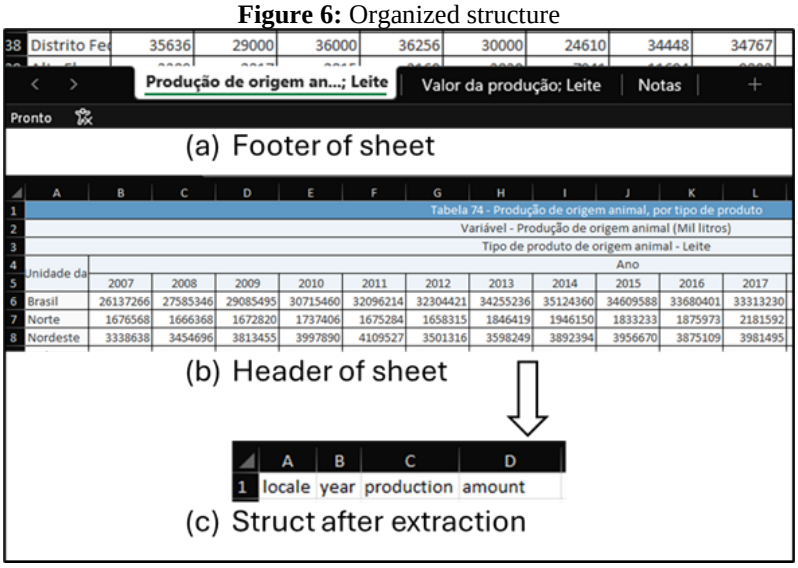


Figure 6: Organized structure



The partial structure that would be used for this case was the generation of a partial CSV file, which would collect data on the location, region belonging to, year, total dairy herd, males up to 6 months, females up to 6 months, males between 7 and 12 months, females between 7 and 12 months, males between 13 and 24 months, females between 13 and 24 months, males between 25 and 36 months, females between 25 and 36 months, males over 36 months, females over 36 months, total males, total females.

After the extraction process, we obtained four partial files, which were then processed and consolidated. It was necessary to standardize the fields, such as the city name, as shown in Figure 8 (a) in the case of the city "Itapuã do Oeste", which had three representations: "Itapuã D'Oeste", "Itapuã do Oeste" and "Itapuã do Oeste (RO)". The first step was to remove the suffix "(RO)" from all the cities that had this representation. The second step was to check the cities with different spellings and standardize them, either by complementing the city name or by standardizing the way it was written. Figure 8 (b) shows part of the function used for this standardization.

Figure 7: Shows the table formatting

INFORME SEMESTRAL DE CAMPO REFERENTE A 23ª ETAPA DE VACINAÇÃO CONTRA FEBRE AFTOSA
ESTADO: RONDÔNIA **BOVINOS DE LEITE**

MUNICÍPIOS	Área Implan-tada Km2.	Rebanho total (bv+bu) existente	Rebanho bovino de leite existente	Percentual em relação ao reb. Total	Bovinos Vacinados por Faixa Etária em Meses									
					Até 8		9 a 12		13 a 24		25 a 36		> de 36	
					M	F	M	F	M	F	M	F	M	
REGIONAL PORTO VELHO														
Porto Velho	34.082,366	568.176	35.089	6,18	2.663	3.135	1.661	1.582	1.708	3.085	1.924	4.252	1.383	
Guajará Mirim	24.855,652	123.017	14.106	11,47	1.046	947	673	962	1.022	1.999	417	1.630	337	
Itapuã D'Oeste	4.081,433	75.621	14.245	18,84	1.234	1.226	710	909	872	1.720	421	1.438	346	
Nova Mamoré	10.071,702	343.179	133.869	39,01	10.608	11.805	6.519	8.249	8.375	12.229	6.773	17.567	3.189	
Candeias do Jamari	6.843,866	164.299	20.536	12,50	1.513	1.519	1.210	1.387	1.328	1.974	1.009	1.972	864	
SUBTOTAL	79.935,02	1.274.292	217.845	17,10	17.064	18.632	10.773	13.089	13.305	21.007	10.544	26.859	6.119	
REGIONAL ARIQUEMES														
Ariquemes	4.426,558	427.798	50.405	11,78	3.664	3.800	2.675	3.292	3.206	4.223	2.776	5.187	2.666	
Alto Paraíso	2.651,818	200.064	24.755	12,37	1.701	2.400	1.481	1.015	1.477	1.775	1.522	3.586	979	
Buritis	3.265,739	344.437	95.382	27,69	7.069	7.663	5.032	6.127	4.768	10.268	3.293	9.821	4.099	

Figure 8: Shows part of the function used for this standardization

Porto Velho	Porto Velho (RO)
Guajará Mirim	Guajará-Mirim (RO)
Itapuã D'Oeste	Itapuã do Oeste (RO)
Nova Mamoré	Nova Mamoré (RO)
Candeias do Jamari	Candeias do Jamari (RO)

(a) Names unstandardized

```
def standard_city(city):
    if city == "Vale Paraiso":
        return "Vale do Paraiso"
    if city == "Guajará Mirim":
        return "Guajará-Mirim"
    if city == "Itapuã D'Oeste":
        return "Itapuã do Oeste"
    if city == "Campo Novo":
        return "Campo Novo de Rondônia"
    if city == "Machadinho":
        return "Machadinho D'Oeste"
    if city in ["Gov. Jorge Teixeira", "Jorge Teixeira"]:
        return "Governador Jorge Teixeira"
```

(b) Transform function

The next necessary transformation was the reduction of the herd age columns so that the animal unit could be calculated. In this case, the animal unit is the measurement that allows the animal load to be estimated, enabling comparison between animals of different ages and categories. The reduction in this case is since the calculation of the animal unit does not differentiate between males and females up to 36 months. Therefore, it was necessary to go from males up to 6 months, females up to 6 months, males between 7 and 12 months, females between 7 and 12 months, males between 13 and 24 months, females between 13 and 24 months, males between 25 and 36 months, females between 25 and 36 months, males over 36 months, females over 36 months, to herd up to 12 months, herd between 13 and 24 months, herd between 25 and 36 months, male herd over 36 months, female herd over 36 months, in addition to creating another category of "Cows in Production", which is an estimate calculated from the sum of the total number of females over 36 months and 70% of females between 24 and 36 months of age[8]. To calculate the animal unit, equation presented in[9] was used as a basis and using the equivalence values presented by the Executive Committee of the Cocoa Farming Plan – CEPLAC[10]. Consider Ha = herd up to 12 months, Hb = herd between 12 and 24 months, Hc = herd between 25 and 36 months and AUt = Total Animal Unit.

$$AUt = \frac{Ha * 106{,}75 + Hb * 208{,}6 + Hc * 293{,}5 + Bulls * 463 + Cows * 450}{450}$$

The calculated columns for birth rate and stocking rate were added. The birth rate is calculated from the relationship between the number of calves, herd up to 12 months, and the number of females of reproductive age[8]. While the stocking rate is calculated from the relationship between the Total Animal Unit and the pasture area. In addition, the metrics calculated from the averages were added:

1. Annual milk production per cow;
2. Value (R$) generated per cow annually;
3. Milk production per cow per day;
4. Value (R$) generated per cow per day;
5. Average value of a liter of milk.

Finally, for the loading phase, the MySQL database was used, which resulted in a table with 24 columns, currently with 884 rows. To visualize the data, Metabase was used, a tool that made it possible to perform visual analyses and create dashboards that are available through the platform "Observatório do Leite de Rondônia" < https://observatoriodoleite.ifro.edu.br/>. This allows for dynamic visualization, as shown in

Figure ***9*** and Figure 10.

Figure 9: Dynamic visualization

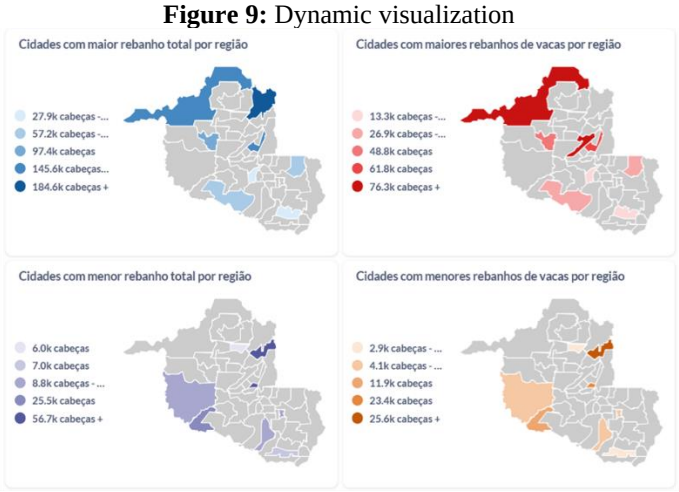
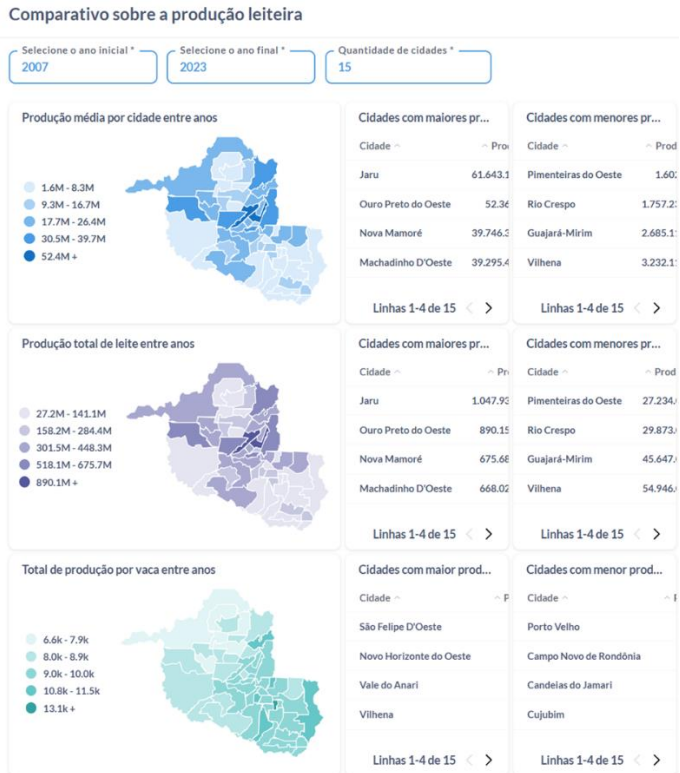


Figure 10: Dynamic visualization



IV. Conclusion

The implementation of the Data Warehouse (DW) for the milk production chain in Rondônia allows the integration and analysis of data from different sources, in addition to easily obtaining transformations of previously uncollected data. This process allows a dynamic analysis of the milk production chain in Rondônia, a production that has been decreasing every year.

Therefore, it is crucial for government agencies to have tools available that allow for user-friendly and dynamic visualization, enabling the verification of scenarios that would not otherwise be possible, or that would require technical knowledge and time to perform.

Therefore, the process of consolidating this DW is strategic from the point of view of building public policies, in addition to enabling its expansion, adding other data sources, such as data from the Municipal Agriculture Survey, Plant Extraction and Forestry Production Survey, and Industrial Sector Surveys. Therefore, with the expansion of the current structure, it will be possible to outline more complex and realistic future scenarios, since currently, all the sources necessary to carry out this study are decoupled, thus it is expected that in the future the tool will prove to be a structuring point for the production chain of the State of Rondônia.

References

[1]. Fao. Dairy Market Review: Overview Of Global Market Developments In 2024. Rome. 2024. From: Https://Openknowledge.Fao.Org/Server/Api/Core/Bitstreams/44217723-F8e7-431a-B3b2-47f10f8fbc71/Content

[2]. Ibge-Ppm, Instituto Brasileiro De Geografia E Estatística. Pesquisa Da Pecuária Municipal. 2024. From: Https://Www.Ibge.Gov.Br/Estatisticas/Economicas/Agricultura-E-Pecuaria/9107-Producao-Da-Pecuaria-Municipal.Html?=&T=O-Que-E

[3]. Hevner A. R., March S. T., Park J. And Ram S. Design Science In Information Systems Research. Mis Quarterly. 2004. V. 28, N. 1, P. 75-105.

[4]. Peffers K., Tuunanen T., Rothenberger M. A. And Chatterjee S. A Design Science Research Methodology For Information Systems Research. Journal Of Management Information Systems. 2008. V. 24, N. 3, P. 45-77.

[5]. Ibge-Cidades, Instituto Brasileiro De Geografia E Estatística. Cidades E Estados Do Brasil. 2024. From: Https://Cidades.Ibge.Gov.Br/

[6]. Universidade Federal De Goiás. Atlas Das Pastagens Brasileiras. 2024. From: Https://Atlasdaspastagens.Ufg.Br/

[7]. Idaron, Agência De Defesa Sanitária Agrosilvopastoril Do Estado De Rondônia. Dados Agropecuários. 2024. From: Http://Www.Idaron.Ro.Gov.Br/Index.Php/Relatorios-E-Formularios/

[8]. Restle J., Pacheco P. S., Pascoal L. L., Pádua J. T., Moletta J. L., Kellermann A. F. And Leite D. T. Efeito Da Pastagem, Da Produção E Da Composição Do Leite No Desempenho De Bezerros De Diferentes Grupos Genéticos. Revista Brasileira De Zootecnia. 2004. V. 33, N. 3, P. 691-703.

[9]. Souza V. L. De. Avaliação Da Fertilidade Do Rebanho Bovino De Rondônia E O Uso Da Detecção De Cio E Do Gnrh Para Aumentar A Eficiência De Programas De Iatf. 2019. 38 F. Dissertação (Mestrado Em Desenvolvimento Regional E Meio Ambiente) - Programa De Pós-Graduação Em Desenvolvimento Regional E Meio Ambiente (Pgdra), Fundação Universidade Federal De Rondônia (Unir), Porto Velho.

[10]. Comissão Executiva Do Plano Da Lavoura Cacaueira – Ceplac. Pecuária. Online. From: Http://Www.Ceplac.Gov.Br/Servicos/Agricolas/Equivalencia.Htm.